Numeral classifiers and number marking in Indo-Iranian: A phylogenetic approach

Cathcart, Chundra ; Hölzl, Andreas ; Jäger, Gerhard ; Widmer, Paul ; Bickel, Balthasar

Abstract: This paper investigates the origins of sortal numeral classifiers in the Indo-Iranian languages. While these are often assumed to result from contact with non-Indo-European languages, an alternative possibility is that classifiers developed as a response to the rise of optional plural marking. This alternative is in line with the so-called Greenberg-Sanches-Slobin (henceforth GSS) generalization. The GSS generalization holds that the presence of sortal numeral classifiers across languages is negatively correlated with obligatory plural marking on nouns. We assess the extent to which Indo-Iranian classifier development is influenced by loosening of restrictions on plural marking using a sample of 65 languages and a Bayesian phylogenetic model, inferring posterior distributions over evolutionary transition rates between typological states and using these rates to reconstruct the history of classifiers and number marking throughout Indo-Iranian, constrained by historically attested states. We find broad support for a diachronically oriented construal of the GSS generalization, but find no evidence for a strong bias against the synchronic co-occurrence of classifiers and obligatory plural marking. Inspection of the most likely diachronic trajectories in individual lineages in the tree shows a stronger effect of the GSS among Iranian languages than Indo-Aryan languages. Taken as a whole, these findings suggest that the association of classifiers and optional number marking in Indo-Iranian is neither solely the effect of universal mechanisms nor of the contingency of local contact histories.

DOI: https://doi.org/10.1163/22105832-bja10013

BRILL

LANGUAGE DYNAMICS AND CHANGE (2020) 1–53

brill.com/ldc

Numeral classifiers and number marking in Indo-Iranian

A phylogenetic approach

Chundra A. Cathcart
Department of Comparative Language Science, and Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Zurich, Switzerland
chundra.cathcart@uzh.ch

Andreas Hölzl
Department of Comparative Language Science, and Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Zurich, Switzerland
andreas.hoelzl@uzh.ch

Gerhard Jäger
Department of Linguistics, University of Tübingen, Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Paul Widmer
Department of Comparative Language Science, and Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Zurich, Switzerland
paul.widmer@uzh.ch

Balthasar Bickel
Department of Comparative Language Science, and Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Zurich, Switzerland
balthasar.bickel@uzh.ch

 | DOI:10.1163/22105832-bja10013

Abstract

This paper investigates the origins of sortal numeral classifiers in the Indo-Iranian languages. While these are often assumed to result from contact with non-Indo-European languages, an alternative possibility is that classifiers developed as a response to the rise of optional plural marking. This alternative is in line with the so-called Greenberg-Sanches-Slobin (henceforth GSS) generalization. The GSS generalization holds that the presence of sortal numeral classifiers across languages is negatively correlated with obligatory plural marking on nouns. We assess the extent to which Indo-Iranian classifier development is influenced by loosening of restrictions on plural marking using a sample of 65 languages and a Bayesian phylogenetic model, inferring posterior distributions over evolutionary transition rates between typological states and using these rates to reconstruct the history of classifiers and number marking throughout Indo-Iranian, constrained by historically attested states. We find broad support for a diachronically oriented construal of the GSS generalization, but find no evidence for a strong bias against the synchronic co-occurrence of classifiers and obligatory plural marking. Inspection of the most likely diachronic trajectories in individual lineages in the tree shows a stronger effect of the GSS among Iranian languages than Indo-Aryan languages. Taken as a whole, these findings suggest that the association of classifiers and optional number marking in Indo-Iranian is neither solely the effect of universal mechanisms nor of the contingency of local contact histories.

Keywords

evolutionary linguistics – historical linguistics – Indo-Iranian – numeral classifiers – plural marking

1 Introduction

Indo-Iranian languages display considerable diversity in constructions where items are enumerated. Ancient languages such as Sanskrit and Avestan, which possess rich nominal morphology, show a straightforward pattern where head nouns agree in number with numerals and quantifiers; however, not all contemporary languages consistently mark plural number on semantically plural nouns. Additionally, several modern Indo-Iranian languages make use of sortal numeral classifiers, as in the Bengali example *chɔ-ṭa boi* 'six books' (six-CLF book), where *ṭa* is a classifying element that co-occurs with an enumerated entity and where the noun does not inflect for number.

Numeral classifiers are typologically uncharacteristic of the larger Indo-European family, and their occurrence within Indo-Iranian has therefore been attributed by some researchers to contact with languages from other stocks (Emeneau 1956, 1965 [1980], Matisoff 1978, Thomason & Kaufman 1988). Others see Indo-Iranian numeral classifiers as the grammaticalized outcome of a general tendency in ancient and medieval Indo-European languages to place generic and non-generic nouns in close apposition, a pattern that can potentially lead to systems of nominal classification (Hackstein 2010). An important confound in resolving this debate is the extent to which Indo-Iranian languages are subject to what is known as the Greenberg-Sanches-Slobin (GSS) generalization (Greenberg 1972, Sanches & Slobin 1973). This generalization posits an association between the presence of numeral classifiers in a language with optionality (or even absence) of plural marking. If numeral classifiers are more likely to develop in languages with optional rather than obligatory number marking, their emergence in Indo-Iranian might reflect a general tendency of the sort captured by the GSS generalization. If not, Indo-Iranian classifiers may have arisen due to contingencies of the family's history, especially its contact history.

This paper explores the diachronic pathways by which the diverse patterns seen in Indo-Iranian have developed. We employ an explicit phylogenetic approach to this question, inferring evolutionary transition rates between typological states concerning the presence of numeral classifiers and optionality of plural marking on the basis of the patterns found in 65 Indo-Iranian languages attested during different chronological periods. First, we use these rates to operationalize two possible interpretations, which we term mutational and selectional, of the GSS generalization. Specifically we observe whether the rate of classifier development is higher in the presence of optional plural marking than in the presence of obligatory plural marking; we also investigate whether co-occurrence between obligatory plural marking and classifier presence is dispreferred in an evolutionary perspective. Subsequently, we use these rates to infer the most likely diachronic trajectories in individual Indo-Iranian phylogenetic lineages, allowing us to identify different pressures in the development of classifiers during the history of the Indo-Iranian languages. While we uncover broad statistical support for the GSS generalization, we find that neither the GSS generalization nor language contact alone can account for all instances of classifier development in Indo-Iranian. Interestingly, the role of optional plural marking in classifier development appears to differ across the two main branches of Indo-Iranian: Indo-Aryan languages develop classifiers less frequently than Iranian languages, but the frequency increases again in the context of contact with languages from non-Indo-European stocks; Iranian lan-

guages, on the other hand, appear largely to have developed classifiers after prolonged periods of optional plural marking, in line with the GSS generalization. Given what is known about the sociopolitical history of the Iranian-speaking area as well as the pattern and matter borrowing (i.e., the transfer of morphosyntactic patterns and phonological material from one language to another; Matras & Sakel 2007) that we observe in our data set, we conclude that Iranian classifiers are largely a response to optional plural marking that was in some cases helped along by contact.

In what follows, we describe the GSS generalization in more detail. We then provide a detailed description of the synchronic and diachronic patterns seen in Indo-Iranian constructions where items are enumerated (Section 2). After introducing our data coding and the model used to test our hypotheses (Section 3), we present and discuss our main findings (Sections 4–6).

2 Background

2.1 *Numeral classifiers*

Numeral classifiers[1] are usually contiguous to numerals in expressions of quantity or, more generally, found to occur in the context of quantification (Grinevald 2000:63). In this study, we define a numeral classifier as any morpheme that, independent of its morphosyntactic status, is linearly adjacent to a numeral (or an equivalent quantifier) when it occurs, and that functions as an attribute of a head noun, together with the numeral. A numeral classifier tends to have mutual dependencies (e.g., collocational, morphosyntactic, phonological, or semantic) with both the numeral and the head noun. In the following example of Mandarin Chinese, for instance, the noun *shū* 'book' can only be combined with the numeral classifier *běn*. Furthermore, the numeral *liǎng* replaces the general numeral *èr* 'two' in numeral classifier constructions. If *běn* is not preceded by a numeral it has a set of different meanings.

(1) Mandarin (Sino-Tibetan)
liǎng běn shū
two.CLF CLF book
'two books'

1 Alternative names such as NUMERATIVE (Aikhenvald 2000:98) and NOMIFIER (Haspelmath 2018) have been proposed for this phenomenon and related phenomena.

Numeral classifiers are generally divided into mensural and sortal subtypes. Mensural classifiers aid in partitioning an uncountable noun (e.g., Modern German *ein Glas Bier* 'a glass of beer').[2] Sortal classifiers, by contrast, are not limited to uncountable nouns, and have been defined variously as:

– A member of a paradigm that forms a binary phrase with a numeral, which in turn forms a binary phrase with a counted noun (Lehmann 2000)
– A grammatical element that occurs with nouns (regardless of their degree of countability) in construction with numerals (Gil 2013)
– An expression that indicates a unit of counting or measure (Doetjes 2012)

Unlike mensural classifiers, sortal classifiers cannot be modified by an adjective: expressions like *ein kaltes Glas Bier* 'a cold glass of beer' have no equivalent among sortal classifiers; sortal classifiers generally cannot co-occur with mensural classifiers, e.g., Maithili *das*(*-*ṭā*) *kap cāy* 'ten (*CLF) cup tea' (Burghart 1992:I:117).

A third type of numeral classifier designates groups, similar to English *a flock of birds* (e.g., Beckwith 1998:131–133). Such classifiers indicate a set larger than one, including a pair. Sortal and mensural numeral classifiers do not indicate any number on their own, but are differentiated by the type of noun they occur with. Most languages exhibit numeral classifiers that refer to measures and groups. This study exclusively focuses on sortal numeral classifiers, which are much less common cross-linguistically and exhibit a very specific geographic distribution.

While a numeral classifier can provide an index to inherent semantic properties of the head noun (its use is often dependent on these properties) and can also carry pragmatic meaning, it does not achieve semantic modification in the way that adnominal elements such as adjectives do.[3] Furthermore, anaphoric use is common, in which case the head noun is excluded.

2.1.1 Indo-Iranian classifiers

The diachrony of Indo-Iranian numeral classifier systems is not well understood because of lacunae in the historical record, making it difficult to determine whether they developed as a response to the loosening of restrictions

2 They are usually considered to be distinct from pseudo-partitives, as in Middle High German *ein glas bier-s* 'a glass of beer' (Bauer 2017:33). Unlike pseudo-partitives, sortal numeral classifiers do not trigger special marking of the head noun.

3 In some languages, elements identical to numeral classifiers also occur in bare use with nouns, marking them variously for definiteness or indefiniteness (Simpson et al. 2011). Though these are often referred to as "bare numeral classifiers," they fall outside the scope of our definition.

on plural marking, because of contact, or due to other factors. Attestations of classifiers are largely absent from pre-modern Indo-Aryan languages, e.g., Old Bengali, perhaps due to stigmatization and suppression in literary registers (for discussion, see Barz & Diller 1985:168). It is however possible to trace the development of certain numeral classifiers in the history of Persian, though the material is incomplete and the exact pathway of development is somewhat ambiguous. In the absence of evidence from the historical record, it is in theory possible to infer whether classifiers developed due to language contact on the basis of (1) the presence of matter borrowing and (2) languages' proximity to groups with numeral classifiers, but here, the picture is not entirely clear.

Lexically speaking, Indo-Iranian classifiers are a mix of inherited and borrowed material. Agia Varvara Romani has borrowed the classifier *-tane* from Turkish[4] (and has obligatory plural marking):

(2) *Dikhlém pándžtane raklá*
see.1SG.PST 5.CLF girl.PL.ACC
'I saw five girls' (Igla 1996:45)

The classifiers found in Indo-Aryan languages of South Asia tend to belong to a core group of elements with transparent Old Indo-Aryan (OIA) etymologies, which are supplemented with additional classifiers (Assamese has roughly a dozen numeral classifiers, all of which appear to be from inherited Indo-Aryan material). The most geographically widespread Indo-Aryan classifier is a reflex of Old Indo-Aryan *jana-* 'person', occurring in Sinhala as the element *denaa* (Geiger 1942:4) as well as in Eastern Indo-Aryan languages. Another common classifier (Nepali *vaṭā*; Bengali, Oriya, *ṭā*; Assamese *tā*) is derived by Chatterji (1926:684 ff.) from OIA *vṛt-ti-ka-*, a deverbal noun built to the root *vart-* 'turn'. The classifier *goṭ*, found in Maithili and other Eastern Indo-Aryan languages, continues Old Indo-Aryan *gōṭṭa-* *'something round' (Turner 1962–1966:229). With the exception of Sinhala, the aforementioned languages have optional plural marking on most noun types, and tend to prohibit plural marking on enumerated nouns that co-occur with numeral classifiers:

(3) Maithili
sāt -ṭā murgī (*-*sabh*)
seven CLF hen PL
'seven hens' (Burghart 1992:117)

4 The Turkish form is generally thought to be an Iranian loan cognate to Persian *dānah* 'grain', with devoicing of initial *d-* (Stilo 2018:138).

However, in Nepali, where optional plural marking is the default, referential scales may require plural marking in some circumstances, leading to the co-occurrence of numeral classifiers and plural marking:

(4) Nepali
cār janā mitra *(*-haru*)
4 CLF friend PL
'four friends' (Bhim Lal Gautam, p.c.)

In other contexts, e.g., for inanimates, Nepali number marking is entirely optional in the presence of classifiers.

Iranian languages employ a more varied mix of inherited forms and borrowed elements from Arabic as well as Turkic languages. Modern Persian employs a number of Arabic terms in addition to the inherited classifier *tā*. It seems unlikely that Persian borrowed these items as classifiers *per se*, since the syntactic patterns characterizing Persian classifier use differ from those of Arabic dialects with numeral classifiers.[5] These classifiers of Arabic origin can be found in related Iranian languages. Classifiers of Turkic origin are found as well. Zazaki *teney* is a Turkish loan (the same element is found in Agia Varvara Romani), and Pashto *tana/teni* may be from a Turkic source as well, though these forms are ultimately Iranian back loans. The Sariqoli classifier *tol* may be of Turkic origin (cf. Uygur *tal*).

The full range of circumstances under which Iranian classifiers developed is unknown, but Middle and Early Modern Persian provide a window onto the usage of the precursor of the Modern Persian general classifier *tā*, which continues Middle Persian (Pahlavi) *tāg*,[6] a multifunctional element glossed as 'item, unit, alone, single' by MacKenzie (1971), who keeps this headword separate from *tāg* 'branch'. The following examples from the Pahlavi Widēwdād show a diverse range of uses:

(5) *nay ēw tāg*
reed one piece
'a single piece of reed' (Moazami 2014:272)

5 For instance, Persian employs a numeral classifier *ra's* 'head' (< Arabic) for animals. Greenberg (1972:18–20) cites constructions from a 19th century Arabic dialect of Oman and Zanzibar which employs the same word when counting animals, but in contrast to Persian usage, it inflects for number.

6 To our knowledge, this is the only classifier-like element in Persian with historical attestation.

(6) *ēw tāg frāz ō ātaxš dahēd*
one piece forward to fire give.IMPER
'give once to the fire' (Moazami 2014:254)

(7) *bōb -ē bālišn se tāg*
fine_carpet INDEF pillow three piece
'a fine carpet (and) three-fold cushions' (Moazami 2014:352)

(8) *spiš -ē ayāb rišk -ē tāg*
louse EZ or nit EZ piece
'a single louse or nit' (Moazami 2014:390)

(9) *u -š nō -īh tāg wēd barīd*
he 3SG nine ABSTRACT.SUFFIX branch willow bring.PST.3SG
'he brought nine branches[7] (of *barsom*)' (Moazami 2014:468)

Mache (2012:171) cites the following example, in which *tā*(*g*) appears to be used as a sortal classifier:

(10) *čand tā dānāgān ī hindūgān*
some tā wise.PL EZ Indian.PL
'some wise Indian men'

She argues on the basis of this example that Middle Persian is a classifier language, though the evidence is quite restricted. We refrain from treating Middle Persian as a numeral classifier language, given the scant, rather ambiguous evidence and the fact that the loose morphosyntactic integration of information in noun phrases containing *tāg* makes it difficult to determine whether they meet the criteria we have chosen for numeral classifiers.

Numeral classifier constructions like those seen in Modern Persian are not attested in Early Modern Persian. However, some constructions involving *tā* 'piece, unit' are found in Early Modern Persian. In these constructions, *tā* marked for indefiniteness is followed by a number or quantifier.[8] Most

7 MacKenzie (1971) keeps the headwords for 'branch' and 'item' separate. Regardless of whether these entries should be separate, the use of 'branch' as a mensural classifier here is noteworthy. We see also the reduplicative plural *tāg tāg* (Moazami 2014:122).

8 Lazard (1963) states that this construction—e.g., indefinite nouns followed by a number or quantifier—indicates an order of magnitude for large numbers ("nombres ronds"), and is used for approximation with small numbers.

instances are limited to one text, the Iskandar-Nāmah, which can be dated to the 12th century CE (Lazard 1963:127), but this usage is conceivably reflective of non-literary usage (Lazard 1963:217–218):

(11) *Zangiyān i nīmkušta tā'ē čand*
Zangi.PL EZ half.dead piece.INDEF few
'some half-dead Zangis (pej. ethnic term)'

(12) *pariyān rā tā'ē saδ bā rasūl bifirist*
Peri.PL OBJ piece.INDEF hundred with messenger send.IMPER
'Send about a hundred Peris, with the messenger'

(13) *tā'ē duvēst rā az Zangiyān bikuštand*
piece.INDEF two.hundred OBJ of Zangi.PL kill.PST.3PL
'they killed two hundred Zangis'

In the above examples, the appositional, loose morphosyntactic integration of elements into the noun phrase is striking, as well as the discontinuity of the noun phrase. In one example, the numeral element is followed by the object marker *rā*, in another, the noun that the numeral element modifies. Adverbial use of *tā'ē* constructions is found as well:

(14) *tā'ē čand bar ān zan zaδ*
piece.INDEF few DAT this woman hit.PST.3SG
'he dealt some blows to this woman'

Additionally, it is worth noting that in the above examples (though they are few in number), *tā'e* constructions co-occur with nouns with overt plural marking, while prenominal numerals co-occur with unmarked nouns. The exact circumstances under which Early Modern Persian constructions came to evolve into Modern Persian numeral classifier constructions remain unclear, if the Early Modern Persian pattern is in fact the diachronic precursor of the Modern Persian one. However, among its multifunctional uses, it is apparent that *tā* serves as an optional means of integrating numeral elements into the noun phrase at both diachronic stages, albeit with differences in the order of the numeral and *tā*, as well as differences in rigidity of the placement of the numeral element with respect to the noun being modified.

It is clear that classifier-like uses of *tāg* were on the rise during Middle Persian times. This is roughly the earliest date at which Turkic and Iranian languages were in contact (Golden 2006). It is possible that Turkic influence

led to the conventionalization in Early Modern Persian of this incipient tendency toward classifier-like constructions, though we have no overt evidence of Turkish influence in the form of matter borrowing (conversely, several Turkic classifiers are made up of borrowed Iranian matter).

Although the Persian historical record provides a slender window onto their development, the conditions that gave rise to numeral classifiers across Indo-Iranian are not entirely clear from the empirical coverage available. We hope to shed further light on the origins of Indo-Iranian numeral classifiers using a probabilistic methodology capable of quantifying the most likely trajectories of classifier development in this subgroup.

2.2 *Optionality of plural marking*

In a given language, individual nouns with plural reference may differ in terms of whether plural number must, cannot, or may be morphologically marked on them. Several Indo-Iranian languages, particularly older ones, have rigorous rules requiring that all semantically plural nouns take plural marking. In some Indo-Iranian languages, phonological and morphological change has resulted in paradigms where morphological plural cannot be marked on some noun types in some cases, i.e., where singular and plural forms are formally identical. In the remaining Indo-Iranian languages, plural marking on some noun types is optional, though it is rarely the case that optional plural marking is allowed on all noun types and pronouns; it is usually required on first person pronouns, at the very least, and tends to be sensitive to referential scales such as the animacy hierarchy (Silverstein 1976). Noun types which take optional plural marking in plural referential contexts are said to exhibit transnumerality or optional plural marking (cf. Corbett 2000:9–19), and it is generally the case that certain kind-denoting nouns can be partitioned into entities via strategies other than plural marking, rendering plural marking on such nouns as optional at best, if even allowed.

In the *World Atlas of Language Structures*, Haspelmath (2013) establishes three degrees of plural optionality (impossible, optional and obligatory), cross-classified against animacy types. In this coding scheme, a language is coded as having obligatory number marking even when the distinction between singular and plural is neutralized in the context of numerals and other quantity words. Since here we are interested in the interaction between numerals, classifiers and number marking we opt for a coding scheme that keeps these three dimensions distinct. Specifically, we code any variation in nominal number marking as optional marking, regardless of whether there is a concomitant numeral (or numeral and classifier) in the same noun phrase. For present purposes we also gloss over distinctions in animacy contexts, or any

other semantic or pragmatic dimension that might regulate the appearance of specific number markers.

2.2.1 Number marking in Indo-Iranian

Full information regarding the Indo-Iranian languages surveyed in this paper can be found in the Appendix. Here, we give a synopsis of the behavior seen across Indo-Iranian in the domain of enumeration, taking into account all attested chronological stages, highlighting examples which we believe to be important.

The presence of classifier systems in Indo-Iranian languages has attracted a fair amount of interest, particularly in contact linguistics (Thomason & Kaufman 1988:85ff.). Numeral classifiers are concentrated in the east of the Indo-Aryan-speaking region (with some possible exceptions described below); Assamese, the easternmost I-A language, has the largest number, with at least a dozen. The proximity of Indo-Aryan languages with numeral classifier systems to mainland Southeast Asia is conspicuous. Emeneau (1956, 1965 [1980]) identifies numeral classifiers as a marker of the Indian linguistic area, but notes also that Indian numeral classifier systems "look like a western outlier of an area whose centre is in East and Southeast Asia" (Emeneau 1965 [1980]:33). Matisoff (1978:78) states that "it seems obvious that the Nepali and Bengali classifier systems are due to T[ibeto]-B[urman] influence, while other Indic languages far removed from the TB area show no signs of developing classifiers."

Heston (1980:147–148) shows that many features which serve as the basis for establishing India as a linguistic area, among them numeral classifiers, are found in Iranian languages as well. She additionally makes the following claim: "Lacking any contrary evidence, there seems no reason to assume the feature [i.e., numeral classifiers] is absent, rather than under-reported, in other Iranian languages [besides Persian and Pashto]." A survey of the evidence shows that the Central Iranian plateau is indeed a hotbed for formally diverse numeral classifier systems, whereas a handful of East Iranian languages appear to have borrowed classifiers from Dari, Tajik or other dialects of Modern Persian.

These views stand in contrast to Hackstein's 2010 study on nominal classification among the Indo-European languages; for Hackstein, numeral classifiers found in Indo-Iranian languages are the grammaticalized endpoint of a family-wide tendency toward apposition between generic and non-generic nouns, though this account leaves unexplained why the distribution of numeral classifiers within Indo-European is so restricted.

The Old Indo-Iranian languages Sanskrit, Avestan, and Old Persian have a rich morphological case and number system, and mark singular, dual, and

plural number on nouns with the corresponding semantic number.[9] Several Middle Indo-Iranian languages show this behavior as well. Pali, the Indo-Aryan language of the Theravada Buddhist Canon, generally maintains a clear morphological distinction between singular and plural; although the nominative singular and plural of *ā*-stems fell together due to regular sound change, a secondary plural suffix was recruited as number marking between the two numbers (Oberlies 2001:150–151). The Middle Iranian languages Khotanese Saka and Khwarezmian consistently mark plural number on nouns with plural reference. For Khotanese Saka, this is largely due to the preservation of a case and number system similar to that of Old Iranian; for Khwarezmian, this may be due to the fact that in a large subset of nouns, morphologized phonological processes such as palatalization render the singular stem distinct from the plural stem, e.g., *'kwnd* 'finger' vs. *'kwnc-n* 'fingers' (Durkin-Meisterernst 2009).

By contrast, a number of Middle Iranian languages do not consistently mark plural on nouns with plural reference, particularly for enumerated nouns. As seen in the following example, plural marking on Middle Persian nouns is entirely optional, particularly when the noun is modified by a numeral:

(15) *ud čahar-dah dar ud mān panz ud gāh sē*
and 14 door and house 5 and throne 3
'and fourteen doors and houses five and thrones three' (Skjærvø 2009:223)

The same pattern holds for Parthian (Durkin-Meisterernst 2014:271). In late Bactrian, case and number distinctions have been neutralized due to the loss of distinctions between final vowels, resulting in an unmarked form without an ending which may be used with either singular or plural reference, and a marked plural form (Sims-Williams 2007:40). Some Sogdian heavy stem nouns show a form that is identical to the singular in plural contexts, e.g., *aβt paxarē-t* vs. *aβt paxarē* 'seven planets', with overt plural marking found only in the former example (Yoshida 2009:313). It is worth noting that Persian, Parthian, and Bactrian nouns tend, in contrast to those of Khwarezmian, to have plural forms that are a straightforwardly affixal extension of a singular "base" form, with no stem alternation (this is true as well for Sogdian heavy stems, in nominative case); this property has been associated with the presence of optional plural marking (Acquaviva 2004:352–354), as there is no overt element that marks singular nouns as unambiguously singular. Old Indo-Iranian languages, in contrast, tended to have plural marking involving more than simply adding

9 Isolated Vedic forms show singular number on nouns with plural reference (Oldenberg 1909).

an affix to the singular form, e.g., Sanskrit *dev-a-ḥ* 'god (nom.sg.)' vs. *dev-ā-ḥ* 'god (nom.pl.)'. In these languages, historical phonological and morphological changes affecting final syllables often resulted in formally identical singular and plural forms, with optional plural marking carried out by suffixes that were previously collective (e.g., Middle Persian *-ān*, *-hā*; Sogdian *-t*), or somehow yielded a similar extensional, straightforwardly affixal pattern.

Modern Indo-Iranian languages show several different patterns. For some languages, plural marking is obligatory. Certain languages of northern Pakistan such as Palula, Kalam Kohistani, and Dumaki consistently mark plural number on nouns with plural reference. Number marking is obligatory in Sinhala, which contains a complex and opaque system comprising at least three noun types: those where the singular and plural are derived from a common base, those where the plural is derived from an unmarked singular, and those where the singular is derived from an unmarked plural. This system appears to have come about via a complex series of developments: initially, plural suffixes were lost in some nouns, leading to a state of affairs where singular marking was optional (Nitz & Nordhoff 2010). In non-enumerative contexts, Ossetic consistently marks plural nouns with the suffix *-t-*, cognate to the Sogdian plural suffix; when a noun is enumerated by a numeral greater than one, the noun is marked by the suffix *-i* (Digor)/*-ɨ* (Iron), synchronically identical to the genitive suffix. According to Thordarson (2008 [2009]:132), this suffix continues the Old Iranian plural suffix **-ah*. The author links this diachronic behavior to that of Yaghnobi, where nouns are marked for oblique case suffix *-i* (perhaps < **-ah*) when enumerated.

Optional plural marking is found in a large number of contemporary languages, including Modern Persian, Kurdish, Zazaki, Bengali, Maithili, Dhivehi, and others. In line with global expectations, the optionality of plural marking in many of these languages is dependent on referential scales, with plural marking often required on nouns of higher animacy. No Indo-Iranian language in our sample allows optional plural marking on pronouns, although some third person pronouns have no morphological distinction between singular and plural.

Another pattern, widespread in Modern Indo-Aryan, involves noun paradigms with morphological restrictions on the expressibility of plural number. For Hindi consonant-final masculine nouns, the direct singular is formally identical to the direct plural, and distinctions in number can be overtly realized only in oblique case forms. Near-identical restrictions of this sort are found in Panjabi, Sindhi, and adjacent Indo-Aryan languages. Similar morphological restrictions can be found in isolated Iranian languages. In Rakhshani Balochi, nouns can be marked for indefiniteness and singularity via the suffix *-e*, but

otherwise, there is no morphological distinction between singular and plural (Barker 1969:3 ff.). In Sangesari, plural is consistently marked on oblique nouns, but cannot be marked on direct nouns, except for a restricted set of items (Azami & Windfuhr 1972:70 ff.). Space does not permit a full investigation into the forces responsible for the development of optional plural marking in Indo-Iranian, though this will undoubtedly prove to be a valuable research direction.

2.3 *The relationship between numeral classifiers and number marking*

The best-known formulation of the observation that languages with numeral classifiers tend to have no, or at best optional, plural marking on at least some noun types (in constructions with numerals as well as those without them) comes from Greenberg (1972) and Sanches & Slobin (1973). This hypothesis, known as the Greenberg-Sanches-Slobin (henceforth GSS) generalization (permutations of the names may vary from work to work), is borne out by a large number of languages. The statistical support for the hypothesis notwithstanding (Tang and Her 2019), there are quite a few exceptions (e.g., Aikhenvald 2000:100–101) and this is also true of South Asia.

As noted in (4), for example, Nepali requires plural marking in the presence of classifiers in at least some circumstances. Similar patterns hold for the Dravidian language Kurux and the Sino-Tibetan language Belhare:

(16) Kurux (Dravidian)
saːt -jʰan kuke -xadd -ar
seven -CLF girl -child -PL
'seven daughters' (Tirkey 2017:389, shortened)

(17) Belhare (Kiranti)
sip -paŋ maʔi -chi
two CLF person NON.SING[ABS]
'two people' (Bickel 2003:563)

Still, the GSS generalization appears to represent a dominant statistical tendency, and a number of proposals have been put forth to explain why plural marking is optional in many languages with numeral classifiers.

A prominent theory proposes that numeral classifiers help or are needed to enumerate, individuate or partition kind-denoting nouns, i.e., nouns like WATER or RICE that involve non-individuated and uncountable reference. The theory furthermore proposes that languages differ in their proportion of kind-denoting nouns (Quine 1960, Silverstein 1976, Lucy 1992, Croft 1994, Krifka 1995). The statistical version of the GSS generalization follows from these two

proposals: languages with more kind-denoting nouns are expected to be more likely to use numeral classifiers in the service of enumeration; furthermore, since kind-denoting nouns are inherently uncountable, number marking is expected to be absent or at best optional on them. Most versions of this theory assume that the proportion of kind-denoting nouns is constrained by a referential scale, e.g., with human-denoting nouns being less likely to be kind-denoting than, say, food-denoting types (Lucy 1992). Theories differ, however, as to whether the variation involves ontological or merely lexical aspects. The ontological view argues that noun types differ cross-linguistically in their ontological entailments: unlike count nouns, kind-denoting nouns designate masses and material without attention to shape and form, and this has ramifications for cognitive domains beyond language (Cassirer 1923, Lucy 1992, Imai & Gentner 1997). Under the lexical view, noun types vary cross-linguistically according to whether the distinction between kind and entity is specified in the lexicon or whether it is lexically ambiguous (Bisang 2002, 2017).

An alternative theory derives an absolute version of the Greenberg-Sanches-Slobin generalization from universal structural configurations. Thus, generative accounts hold that classifiers and plural markers occupy the same structural position (Borer 2005). This predicts the incompatibility of numeral classifiers and plural marking. When they do co-occur nevertheless, as in Nepali, Kurux or Belhare, the relevant markers are predicted to differ from those in other languages, either because of different formal properties (e.g., the classifiers might not be real classifiers) or independent surface phenomena (e.g., the number marker appears where it does for phonological, not syntactic reasons).[10]

At the same time, other theories reject the GSS generalizations and derive the presence of numeral classifiers from properties that are not related to number marking: Aikhenvald & Dixon (1998) derive the probability of numeral classifiers from a general typological variable of reference classification. Gil (2013) sees them as an arbitrary conventionalization, possibly related to less configurational noun phrases (Gil 1987). Lehmann (2010) argues that in some languages they are simply necessary to give a numeral the status of a full word. Under all these theories, the distributions of numeral classifiers and optional number marking reflect independent historical contingencies, especially effects of language contact.

10 The generative account provided by Gebhardt (2018) argues that Persian *tā*, which can co-occur with nouns marked for plural number, is not a classifier in the sense that classifiers in languages such as Chinese are.

In what follows we assess the GSS generalization empirically, probing the evidence for or against a diachronic correlation between numeral classifiers and optional number marking. While overall the Indo-Iranian data seem to be in line with the correlation, a number of observations challenge it:

– Optional plural marking exists in a number of Indo-Iranian languages, likely a diachronic consequence of the loss of final syllable nuclei, and many of these show no sign of developing numeral classifiers. More generally, optionality of plural marking is not a strong predictor for the development of numeral classifiers in Indo-European. Some IE languages (e.g., Hittite) have optional plural on nouns with numerals only, and nowhere else; in some IE languages (e.g., Breton), singular number is even compulsory with numerals. None of these languages, however, developed classifiers.
– Modern Persian requires plural marking in certain referential contexts, in which case it can co-occur with numeral classifiers, flying in the face of the apparent incompatibility of overt plural marking and classifier use.
– A not insignificant number of Indo-Iranian languages, such as Nepali, Kumzari, Yaghnobi, Pashto, and Sinhala, have numeral classifiers and obligatory plural marking.

These observations raise serious questions as to whether the Indo-Iranian patterns owe to development of general number and a subsequent need to partition kind-denoting nouns. In view of this, we turn to statistical modeling to assess the hypothesis. From the predictions enumerated above, we define two versions of the GSS generalization that can be tested using a phylogenetic model. The MUTATIONAL GSS generalization predicts that languages develop numeral classifiers with higher frequency in the presence of optional plural marking than in the presence of obligatory plural marking, since classifiers aid in partitioning kind-denoting nouns. The SELECTIONAL GSS generalization holds there is something inherently dispreferred about the structural combination of classifiers and obligatory plural marking, and that this co-occurrence will be overall less diachronically stable than that of classifiers and optional plural marking. The selectional/mutational dichotomy we have established here can be compared to the distinction between source-oriented and result-oriented explanations drawn by a number of scholars (e.g., Cristofaro 2012; see Schmidtke-Bode 2019 for a survey). We describe our operationalization of these hypotheses in further detail below.

3 Data and methods

We employ an explicitly phylogenetic method to address the mutational and selectional versions of the GSS generalization, as described above. Given a

phylogenetic representation of the languages in our sample, and assuming that change in the linguistic features studied in this paper can be modeled according to a continuous-time Markov (CTM) process, we can quantify and approximate the temporal rates at which transitions between different feature variants occur. These rates can be used to operationalize a wide range of questions regarding the diachronic dynamics of the features under study, and can be used to reconstruct probable trajectories of their development. We use the rates themselves to quantify the overall strength of the GSS generalization over the tree, while stochastic character mapping allows us to disaggregate this information and explore individual languages' histories, pinpointing developments that may be due to contact and other factors not explicitly addressed by our methodology.

3.1 *Data*

The phylogenetic comparative methodology that we employ involves two key ingredients: a tree sample of 1000 phylogenies of the languages under investigation, and a featural representation of each language of interest.

3.1.1 Tree sample

We infer timed phylogenetic trees for Indo-Iranian languages in our sample on the basis of characters automatically extracted from ASJP data (Wichmann et al. 2018) according to the method described in Jäger 2018.[11] We use RevBayes (Höhna et al. 2016) to carry out phylogenetic inference, placing clade constraints on Indo-Aryan and Iranian as well as on uncontroversial subgroups within these clades such as Eastern Indo-Aryan, Insular Indo-Aryan, Southwest Iranian and Sakan (see Masica 1991, Cathcart 2015, Deo 2018 for discussion of these subgroups). Additionally, we constrain the tree such that Old Persian is the ancestor of Middle Persian and Modern Persian dialects and Middle Persian is the ancestor of Modern Persian dialects, an uncontroversial ancestry relationship. Ancient and medieval languages serve as calibration points for our tree, with dates sampled from uniform priors with the parameterizations listed in Table 1; in addition to these calibration points, we place a Uniform(3600, 3800) prior over the root age of the tree; this represents a range of dates (in years before present) directly following the breakup of the Bactriana-Margiana Archeological Complex, accepted by most specialists as the staging ground for Indo-Iranian dispersal (Kuz'mina 2007). We use a

11 We use ASJP's 40-item word lists rather than larger word lists due to the broader language coverage of this resource.

TABLE 1 Calibration points for languages in the tree in years before present; dates are sampled from uniform distributions with the parameters shown

Language		Max age (YBP)	Min age (YBP)
Sanskrit	sans1269	3000	3400
Pali	pali1273	2000	2300
Prakrit	maha1305	2000	2300
Old East Rajasthani	dhun1238	200	0400
Avestan	aves1237	2600	2800
Khotanese Saka	khot1251	1100	1300
Khwarezmian	khwa1238	800	1000
Sogdian	sogd1245	1000	1200
Bactrian	bact1239	1000	1200
Parthian	part1239	1800	2000
Middle Persian	pahl1241	1400	1800
Old Persian	oldp1254	2300	2500
Modern languages		0	250

Birth-Death tree prior and a generalized time-reversible model of trait evolution (Tavaré 1986) with gamma-distributed variation across rate classes and a relaxed clock with log-normal-distributed branch-level rate multipliers. We run 10,000,000 iterations of Markov chain Monte Carlo, discarding the first half of samples and thinning the posterior tree sample to 1000 trees. The tree sample can be seen in Figure 1.

3.1.2 Data coding

We surveyed grammars of 65 ancient, medieval and modern Indo-Iranian languages. The data collected consist of feature codings for each language according to two variables of interest, concerning (1) the presence of numeral classifiers and (2) optionality of and restrictions on plural marking. These variables show the following attested values:

1. Numeral classifiers
   a. Classifiers present: +CLF
   b. Classifiers absent: –CLF
2. Plural marking
   a. Optional plural marking in at least some contexts: –OBL.PL
   b. Obligatory plural marking in all contexts: +OBL.PL

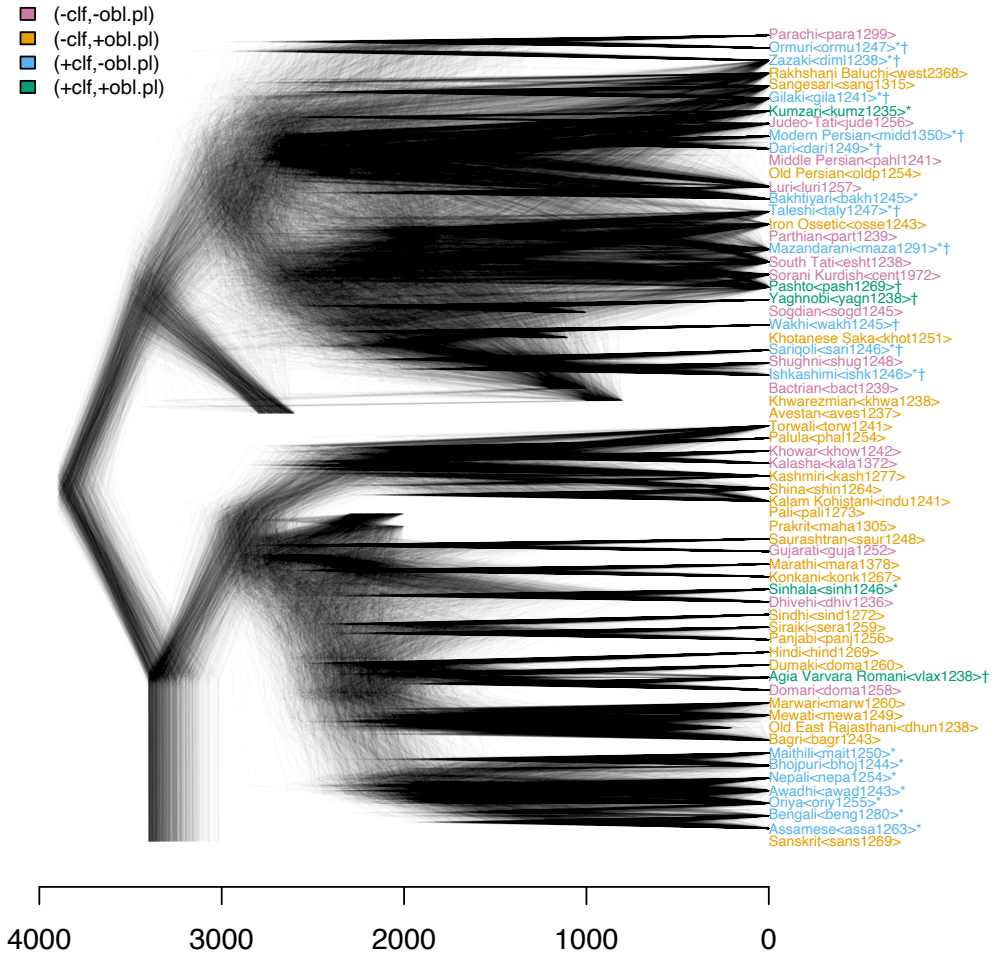

FIGURE 1 Sample of 1000 Indo-Iranian phylogenetic trees; tip colors represent languages' states; for languages with numeral classifiers, * indicates the presence of classifiers based on inherited matter, while † indicates the presence of classifiers based on borrowed matter (but not necessarily borrowed as classifiers *per se*). A maximum clade credibility (MCC) tree can be found in Figure 7. The scale represents age measured in years before present.

Combinations of values for these variables in our sample can be seen in Figures 1 and 2. To ensure that our inference procedure is tractable and meaningful, we keep the number of levels for each variable low. We treat languages with morphological restrictions on plural marking as having obligatory plural marking, since languages with restricted plural marking tend to mark plural number on nouns to the extent possible, whereas languages with optional plural marking choose not to mark plural number in contexts where it is possible (languages with morphologically restricted plural marking are denoted by

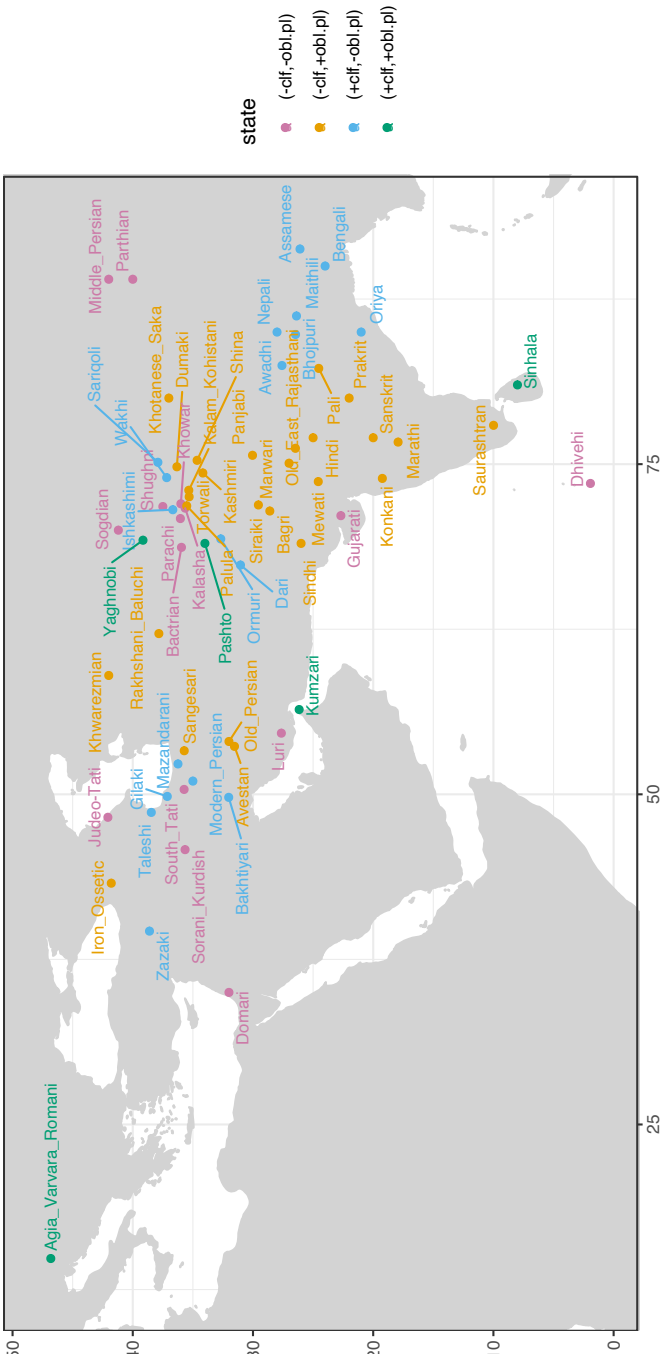

FIGURE 2 Approximate locations of languages in sample, based on closest glottocode matches; ±CLF stands for presence/absence of sortal numeral classifiers, ±OBLPL for presence/absence of obligatory plural on nouns in a language

the state MORPH.PL in the Appendix). This leaves us with a single binary feature ±OBL.PL. Furthermore, we treat all four attested combinations of values of ±CLF and ±OBL.PL as a single feature with multiple values; e.g., Bengali has the value (+CLF, −OBL.PL), as it contains contexts where plural marking is optional. This pared-down representation of the typological state space facilitates ease of phylogenetic inference; we discuss ways of incorporating more fine-grained information in the conclusion section of this paper.

3.2 *Model and inference*

3.2.1 CTM models of character evolution

We model changes between different feature states via a common phylogenetic comparative method, the continuous-time Markov (CTM) process of character (i.e., feature) evolution. Under such a model, transitions between different states (i.e., feature variants) take place at non-negative evolutionary RATES, the inverse of which represents the average time the system spends in a given state. Rates between different states can be found in the off-diagonal cells of the instantaneous rate matrix Q; diagonal cells of the matrix take values such that rows sum to zero.[12] We place prior distributions over the rates in Q such that transitions occur over realistic time intervals, and infer posterior distributions of each rate, as defined below:

$$(18)\quad P(Q|D, \boldsymbol{T}) \propto P(D, Q|\boldsymbol{T}) = \sum_{T \in \boldsymbol{T}} P(D, Q|T)P(T|\boldsymbol{T}) \approx \frac{1}{|\boldsymbol{T}|} \sum_{T \in \boldsymbol{T}} P(D|T, Q)P(Q)$$

D represents the observed linguistic data; $\boldsymbol{T}$ represents the sample of trees. The probability of the data given a tree and set of rates, $P(D|T, Q)$, can be efficiently computed via the Pruning Algorithm (Felsenstein 2004:251–255). Once the posterior distributions of the rates are inferred, they can be used to reconstruct the probability of a given character state at internal nodes of the tree (i.e., nodes where no data are observed). Posterior rates can also be used to carry out STOCHASTIC CHARACTER MAPPING (SCM; Nielsen 2002, Huelsenbeck et al. 2003, Bollback 2006), an iterative process which samples locations on branches of the phylogeny where changes between states have the highest posterior probability of occurring.

12 For a given timespan t, the row-stochastic matrix P_t of transition PROBABILITIES between all states (along with self-transitions) can be computed via matrix exponentiation: $P_t = \exp\{Qt\}$

3.2.2 Phylogenetic hypothesis testing

The literature on the relationship between classifier presence and optionality of plural marking surveyed above makes the prediction that certain pathways of diachronic development will be highly disfavored, if not impossible. The mutational interpretation of the GSS generalization predicts that classifiers will be gained more frequently if the previous state is (–CLF, –OBL.PL) than if the previous state is (–CLF, +OBL.PL). What we term the selectional interpretation predicts that if the state (+CLF, +OBL.PL) is synchronically dispreferred, then the rate at which languages abandon this state will be higher than the rate at which they abandon the state (+CLF, –OBL.PL).

In linguistics, phylogenetic comparative methods provide a means of testing for associations between pairs of linguistic features as they evolve in a tree over time. A standard way of testing for correlated evolution between two discrete binary features, such as ±CLF and ±OBL.PL is Pagel's 1994 Discrete model, which assesses the relative model fit of a dependent model, which constrains evolutionary rates in a manner thought to be compatible with correlated patterns of evolution, against a null, independent model, which models two independent character histories for each of the features in question (Pagel & Meade 2006, Dunn et al. 2011). In the Bayesian context, a common practice is to carry out this assessment using Bayes Factors (i.e., the ratio of marginal likelihoods for each model). In this paper, we largely depart from this framework for the reasons listed below.

First, the Discrete model tells us whether there is support for interdependent evolution, but suppresses most of the dynamics of change over the tree, including directionality of change. Since directionality is built into our hypothesis, we prefer to observe rates from a single model in order to determine whether the classifiers develop more frequently in the presence of optional plural marking—not simply whether a change in one feature correlates with a change in the other feature. Second, the Discrete model has been shown to exhibit problematic behavior under certain circumstances (Maddison & Fitzjohn 2014); in particular, scenarios in which features undergo relatively infrequent changes over the tree can be prone to false detection of the presence of correlated evolution, though this is not a problem for all datasets. Third, while Bayes Factors have traditionally been viewed as a lean way of comparing nested models, statistical science is gradually moving away from their use for the purpose of hypothesis testing in favor of alternative approaches, for both technical and conceptual reasons (Gelman & Shalizi 2013).

Given these concerns, the bulk of our analysis makes use of the computationally simpler approach of carrying out hypothesis testing within a single model (cf. Kruschke 2011) and allowing the most plausible evolutionary story to

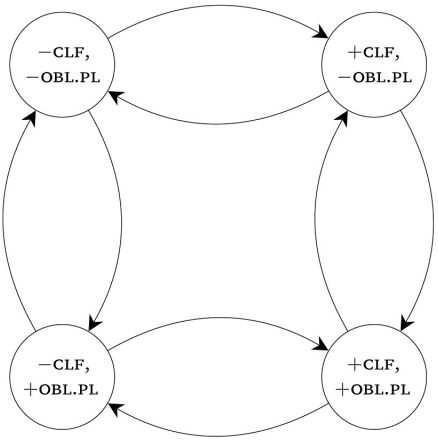

FIGURE 3
Character states and allowed state transitions of the model used in this paper

fall out of these results. Our model involves transition rates between all possible combinations of (±CLF, ±OBL.PL) as long as these rates reflect a single state change, i.e., we rule out the possibility of an instantaneous change from, say, (-CLF, -OBL.PL) to (+CLF, +OBL.PL).[13] The model is schematized in Figure 3. In our analyses, we observe whether credible ranges of posterior values (or quantities derived from these values) are compatible with the hypotheses explored in this paper. Credible intervals (CIs) are widely used in Bayesian statistics, but there is no general consensus on the cutoff that should be employed (McElreath 2020:54ff.). We use a 95% CI, which is conservative in the sense that it leads us to rule out a hypothesis only if it is supported by fewer than 5% of posterior samples. Since most of our hypotheses are one-sided (i.e., is one change type more frequent than another?), we base our assessments on the upper 95% range of posterior values.

3.2.3 Inference

We scale branch lengths by dividing them by 1000, and place a Gamma(1, 1) prior over the rates Q, representing a prior expectation of one change per millennium on average. We sample from the posterior distributions of the evo-

13 This model, which allows only one-step cascading changes, received marginal Bayes Factor support over an unconstrained model where all logically possible transitions were permitted (even those involving simultaneous two-state changes), and a Reversible-Jump model (Pagel & Meade 2006) which allowed certain transition rates to be set to zero. We note that the Bayes Factor is a less-than-ideal way in which to make such a model selection choice, but wish to emphasize that the choice of these three models has no real bearing on the questions under investigation in this paper.

lutionary rates (defined in eq. (18)) using the No U-turn Sampler of `RStan` (Carpenter et al. 2017), aggregating posterior values over trees in the sample. Posterior samples of evolutionary rates can be used to simulate ancestral states at each unobserved node and simulate character histories over the tree, shedding light on likely evolutionary trajectories involving the features of interest.[14]

4 Results

In this section, we assess the overall extent to which Indo-Iranian classifiers have developed in line with the GSS generalization, according to the separate versions of the GSS generalization defined above. In the subsequent section, we analyze individual disaggregated diachronic trajectories. The posterior rates can be seen in Figure 4.

4.1 *Mutational GSS generalization*

The mutational GSS generalization predicts that classifiers are gained more frequently in the presence of optional plural marking than in the presence of obligatory plural marking. We quantify this difference by comparing the posterior rates for the transition $q((-\text{CLF}, -\text{OBL.PL}) \rightarrow (+\text{CLF}, -\text{OBL.PL}))$ with those for the transition $q((-\text{CLF}, +\text{OBL.PL}) \rightarrow (+\text{CLF}, +\text{OBL.PL}))$. Figure 5 gives these rates, as well as the difference between their posterior distributions (i.e., $q((-\text{CLF}, +\text{OBL.PL}) \rightarrow (+\text{CLF}, +\text{OBL.PL}))$ subtracted from $q((-\text{CLF}, -\text{OBL.PL}) \rightarrow (+\text{CLF}, -\text{OBL.PL}))$ for each sample in the posterior trace); positive values indicate a higher preference for classifier gain in the presence of optional plural marking. 98.4% of posterior samples show a positive difference between the two rates. This value, above our 95% CI, indicates substantial support for the GSS generalization in its diachronic form; the development of classifiers in Indo-Iranian languages appears to have been strongly influenced by the presence of optional plural marking. The ratio of the transition rate $q((-\text{CLF}, -\text{OBL.PL}) \rightarrow (+\text{CLF}, -\text{OBL.PL}))$ to the transition rate $q((-\text{CLF}, +\text{OBL.PL}) \rightarrow (+\text{CLF}, +\text{OBL.PL}))$ is greater than one in the majority of posterior samples; this ratio is greater than 1 in 98.4% of samples, greater than 2 in 91.4% of samples, greater than 3 in 81.2% of samples, and greater than 4 in 70% of samples (this ratio is greater than 1.58 in the upper 95% of samples). Despite our conservative cut-

14 Code available at https://github.com/chundrac/clf-iir-evo.

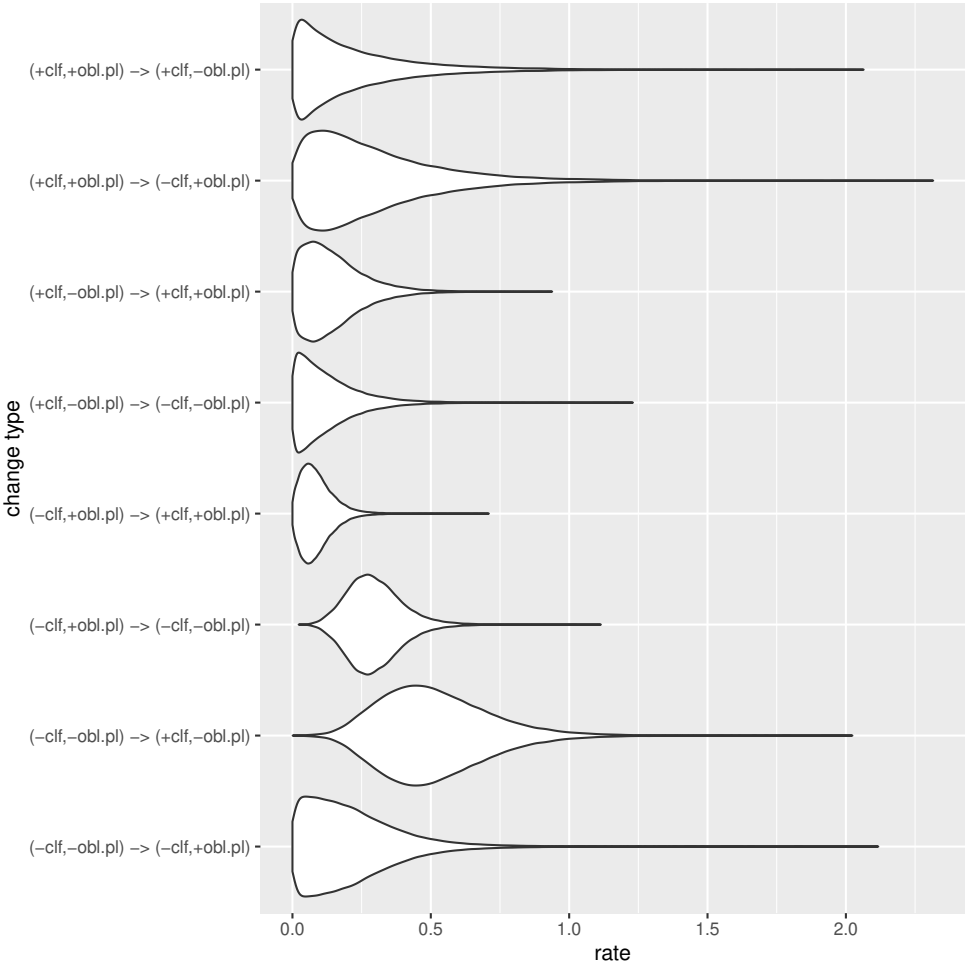

FIGURE 4 Posterior distributions of rates for each change type. Abbreviations as in Figure 2

off for credible ranges, it is still the case that a small proportion of samples go against the GSS hypothesis. For this reason, we investigate branch-specific developments in detail by carrying out STOCHASTIC CHARACTER MAPPING (SCM) in §4.3, which allows us to draw inferences regarding the most probable trajectory leading to the development of classifiers on each branch where they emerge in the tree.

4.2 *Selectional GSS generalization*

The selectional GSS generalization predicts that the state (+CLF, +OBL.PL) is synchronically dispreferred, and we expect this typological state to have a

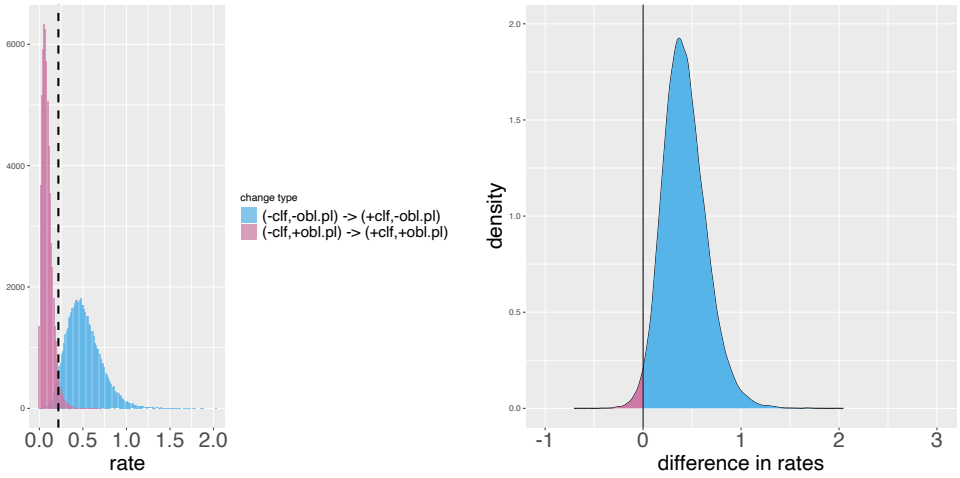

FIGURE 5 Left: posterior distributions of rates for classifier gain given obligatory plural marking versus classifier gain given optional plural marking. Right: Difference between rates of classifier gain in the presence of optional versus obligatory plural marking; values greater than zero indicate that classifiers are gained more frequently in the presence of optional plural marking as opposed to obligatory plural marking.

higher exit rate than the state (+CLF, −OBL.PL). The exit rate of a state can be computed by summing over all transition rates away from the state in question. Hence, the exit rates for the relevant states are the following:

$$q_{\text{exit}}((+\text{CLF}, +\text{OBL.PL})) = q((+\text{CLF}, +\text{OBL.PL}) \to (+\text{CLF}, -\text{OBL.PL})) + q((+\text{CLF}, +\text{OBL.PL}) \to (-\text{CLF}, +\text{OBL.PL}))$$

$$q_{\text{exit}}((+\text{CLF}, -\text{OBL.PL})) = q((+\text{CLF}, -\text{OBL.PL}) \to (+\text{CLF}, +\text{OBL.PL})) + q((+\text{CLF}, -\text{OBL.PL}) \to (-\text{CLF}, -\text{OBL.PL}))$$

Figure 6 reveals that the exit rate for (+CLF, +OBL.PL) is higher than the exit rate for (+CLF, −OBL.PL), but not substantially so; the difference in rates is greater than zero in only 75.6% of samples. The fact that 24.4% of samples are incompatible with the selectional GSS generalization means that we cannot rule out the null hypothesis that (+CLF, +OBL.PL) and (+CLF, −OBL.PL) are roughly equal in their stability, according to our criterion of a 95% CI. This shows that there is nothing inherently dispreferred about the co-occurrence of classifiers and obligatory plural marking; this state of affairs may arise less frequently through diachronic change than the state (+CLF, −OBL.PL): the more frequent trajectory (−CLF, −OBL.PL) → (+CLF, −OBL.PL) may reflect a more gen-

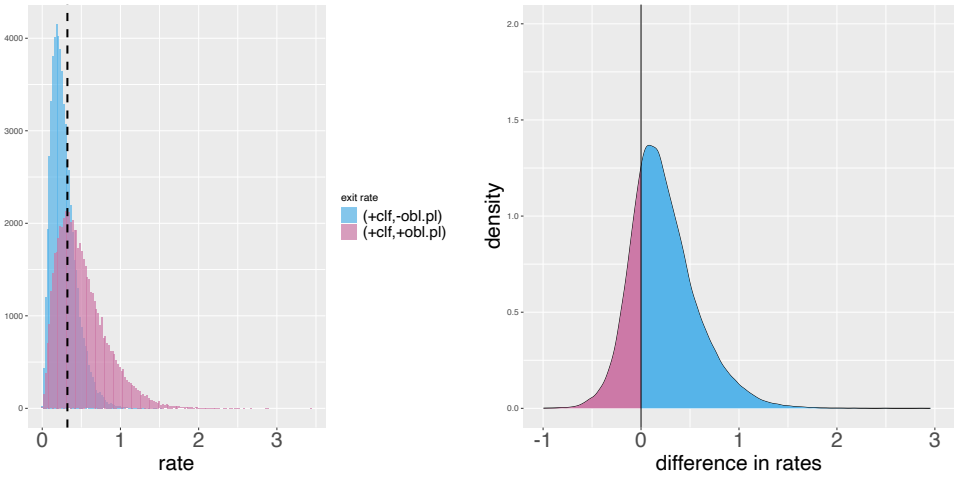

FIGURE 6 Left: posterior distributions of exit rates for the states (+CLF, −OBL.PL) and (+CLF, +OBL.PL). Right: Difference between exit rate for (+CLF, +OBL.PL) and exit rate for (+CLF, −OBL.PL); values greater than zero indicate that (+CLF, +OBL.PL) is abandoned at a higher rate

eral bias or pressure towards unitization, while the trajectory (−CLF, +OBL.PL) → (+CLF, +OBL.PL) may occur for sociolinguistic and contact-based reasons.

Taken together with the results of the mutational GSS generalization, these results indicate that any apparent bias against the state (+CLF, +OBL.PL) that can be detected synchronically across the Indo-Iranian languages (and perhaps beyond) appears to have emerged from diachronic preferences towards the development of certain structures in specific contexts rather than structural constraints on co-occurrence.

4.3 *Character histories*

We carry out stochastic character mapping using the SIMMAP method (Bollback 2006) as implemented in the `R` package `Phytools` (Revell 2012). We simulate character histories on a maximum clade credibility (MCC) tree constructed from our tree sample over 1000 iterations, drawing from the posterior sample of transition rates. The standard way for visualizing the aggregation of these histories is to use a density map, which represents the probability of a state in continuous space over the tree using a color gradient. Visualization can be a challenge for more than two states, since colors can become muddy in regions where uncertainty over the state value of the character is high (Figure 7). For this reason, we supplement this visualization by an alternative one that is based on maximum a posteriori (MAP) estimates (Figure 8). We derive the MAP estimates by tabulating for each branch the counts for each type of

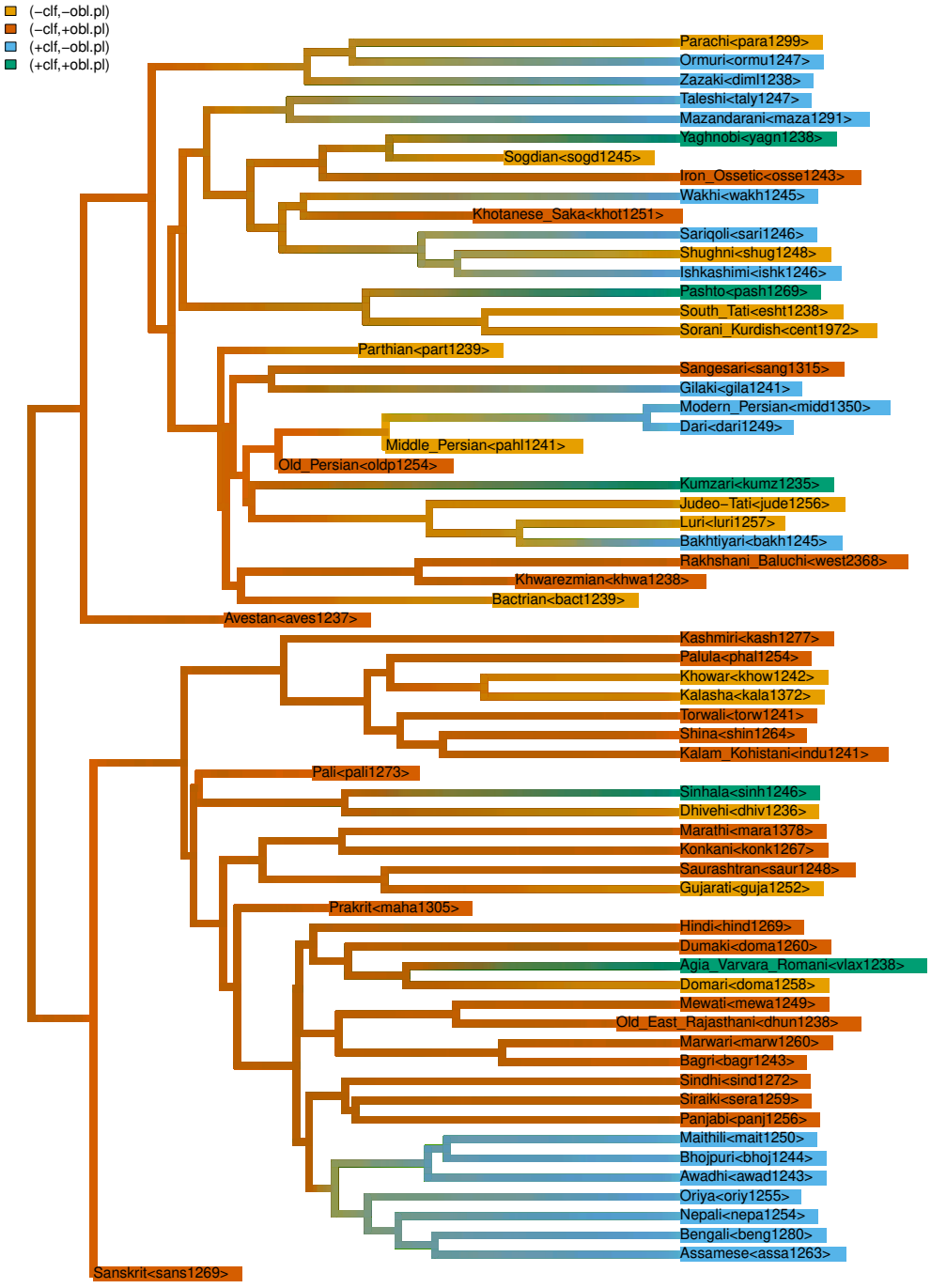

FIGURE 7 Density map aggregating the most probable character histories over a Maximum Clade Credibility (MCC) tree constructed from the tree sample

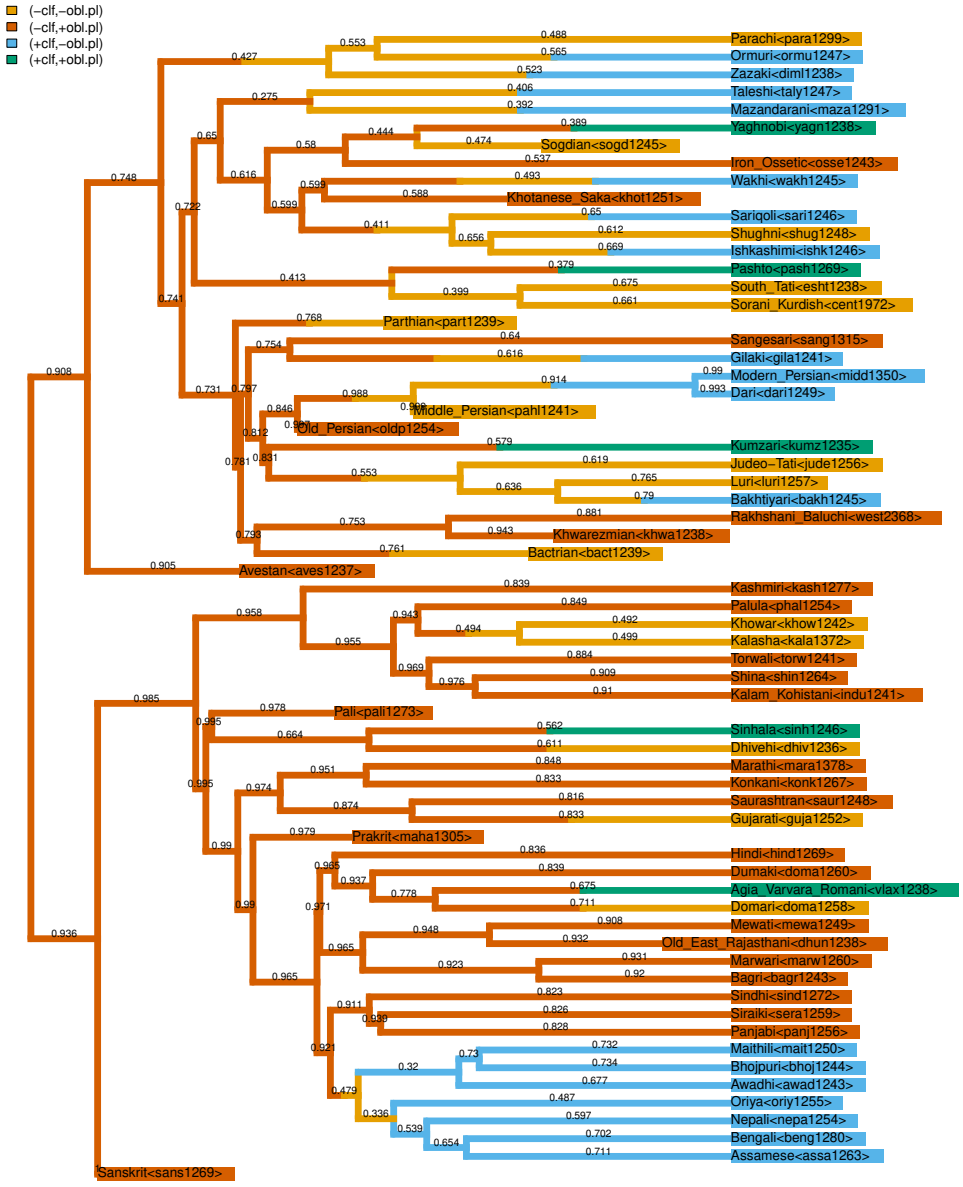

FIGURE 8 Maximum A Posteriori (MAP) character history over a Maximum Clade Credibility (MCC) tree constructed from the tree sample. The numbers in each branch report the posterior support for the MAP transition history.

transition history (ignoring the actual waiting times between transitions) and taking the most frequent transition history.

The figures reveal striking differences in diachronic behavior between Indo-Aryan and Iranian. In Indo-Aryan, classifiers emerge only three times (on branches ancestral to Sinhala, Agia Varvara Romani, and several Eastern Indo-Aryan languages), and their development is not preceded by a long period of optional plural marking. In contrast, an overwhelming number of cases of classifier development in Iranian are preceded by lengthy periods of optional plural marking.

To ensure that these patterns (and specifically, this difference across the two subgroups) are not simply an artifact of the topology of the MCC tree, we carry out SCM on 1000 trees drawn from the tree sample, tabulating the number of times classifiers are gained in the presence of optional plural marking versus obligatory plural marking within Indo-Aryan and Iranian. The results are summarized in Figure 9 and they indicate that for Iranian languages, classifiers develop more frequently in the presence of optional plural marking than in Indo-Aryan languages.

5 Discussion

Taking all the findings together, the mutational version of the GSS generalization has considerable support in the development of classifiers in Indo-Iranian, but the Indo-Aryan results in particular suggest that classifier development is not solely a response to optional plural marking. In what follows, we discuss these developments individually, assessing the role of different factors that potentially underlie the development of numeral classifiers. What is of particular interest in this are potential effects from language contact. To assess these it helps to distinguish between cases in which Indo-Iranian languages have participated in matter borrowing of classifiers as opposed to potential instances of pattern borrowing without any transfer of matter between languages (Matras & Sakel 2007), as indicated by the coding scheme in Figure 1.

5.1 *Indo-Aryan*

As observed above, the MAP tree shows three clear instances of Indo-Aryan classifier development: in Agia Varvara Romani, Sinhala and Eastern Indo-Aryan. In Agia Varvara Romani and Sinhala, classifier development was most likely not preceded by a period of optional plural marking, and in the case of Eastern Indo-Aryan, the period of optional plural marking was relatively short (and perhaps even an artifact of the character model we used, which only

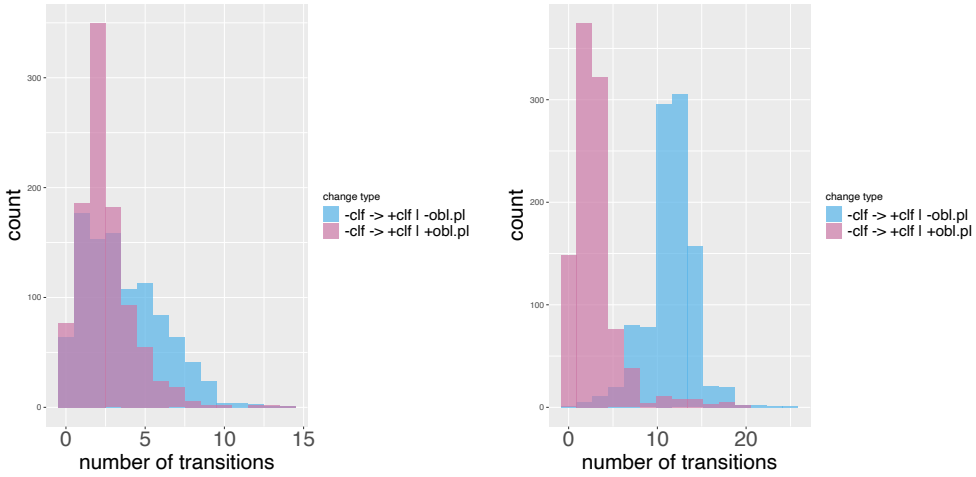

FIGURE 9 Number of gains of numeral classifiers in the presence of optional versus obligatory plural marking for Indo-Aryan (left) and Iranian (right) over 1000 iterations of SCM carried out using trees from the tree sample.

permits transitions involving a single change of state). These patterns favor language contact as a more plausible explanation for the presence of classifiers in Agia Varvara Romani and Sinhala, and perhaps this is even the case in Eastern Indo-Aryan due to the relatively rapid succession of cascading state changes.

Agia Varvara Romani shows clear evidence for matter borrowing of the classifier *tane* from Turkish (see above), possibly due to the prestige of the latter language. The same element is also found in other Romani dialects that are in contact with Turkish (Matras 2002:204). In contrast, the Sinhala classifier consists of inherited phonological material, and therefore does not point explicitly to contact in the form of matter borrowing. Our quantitative results point to a period of obligatory plural marking preceding the emergence of classifiers (MAP support = .53). Incidentally, the Sinhala classifier construction is very similar to that of surrounding Dravidian languages, especially Tamil. Sinhala *denaa* still has its meaning 'people' in other contexts (Chandralal 2010). In Modern Literary Tamil (Lehmann 1993:112–114), numerals above one have a so-called pronominalized form with a suffix *-ar*. In Modern Spoken Tamil (Schiffmann 1999:132–135), pronominalized numerals higher than one add the noun *peeru* 'name' instead, which can also mean 'person'. In both languages the newly formed numeral can be used attributively or pronominally. If used attributively, it can precede or follow the head noun. In Literary Tamil, they have a marked genitive reading instead if preposed. In Spoken Tamil, they have a specific or

definite reading if postposed. The similarities with Sinhala are striking. In all three languages there is [N [Num Clf]] word order as at least one possibility and there is a connection with animacy, non-animate or non-human entities being unmarked. Some Dravidian languages on the mainland have similar classifier-like constructions for humans. In Telugu, for instance, numerals above eight combine with the word *mandi* 'persons', e.g. *padi-mandi* 'ten persons' (Krishnamurti & Gwynn 1985:106–109). This makes pattern borrowing from adjacent Dravidian languages such as Tamil a likely source of the Sinhala classifier construction.

For Eastern Indo-Aryan there is some evidence that medieval varieties had optional plural marking. Mukherjee (1963:23) states that Old Bengali lacks morphological number but has a number of optional periphrastic means of expressing plurality. It is not clear whether this apparent optionality of plural marking co-existed with classifier use, as there is good reason to believe that classifier use was suppressed in literary registers, our only source of data on languages of this sort (cf. Barz & Diller 1985). Regardless of these specific scenarios, our quantitative results suggest that even if these languages went through a period of optional number marking, it was a relatively short period (Figure 8). That such changes took place relatively rapidly suggests a historical development best compatible with widespread language shift.

Support for this interpretation comes from the fact that Eastern Indo-Aryan languages with numeral classifiers are spoken in the vicinity of Dravidian (e.g., Kurux), Austroasiatic (especially Munda, e.g. Kharia), and Sino-Tibetan (especially Tibeto-Burman, e.g., Jero) languages, many of which also exhibit numeral classifiers. Striking parallels with the behavior of East Indo-Aryan classifiers can be found in the Kradai language Khamti. In Khamti, [N Clf Num] word order has an indefinite reading with the numeral 'one', while [N Num Clf] word order is definite. Interestingly, Assamese, Bengali, and Oriya have a connection of word order with definiteness, where [N Num Clf] word order is definite while [Num Clf N] order is indefinite.

(19) Khamti (Kradai)
kuun⁴maau koo¹ leeung³
bachelor CLF one
'a bachelor' (Inglis 2007:8, shortened)

(20) *kuun⁴ saam koo¹*
person three CLF
'three people' (Inglis 2007:11, shortened)

(21) Bengali
chɔ -ṭa boi
six -CLF book
'six books' (David 2015:136)

(22) *boi chɔ -ṭa*
book six -CLF
'the six books' (David 2015:137)

While further parallels and matches in morphosyntactic pattern are needed in order to make a conclusive case for a shift from a language specifically like Khamti to Eastern Indo-Aryan, the striking differences from other Indo-Aryan languages, as well as the dynamics of change shown in the evolutionary scenario that we infer, lend support to the idea that the typological profile of Eastern Indo-Aryan languages is due to contact rather than diachronic trends realized elsewhere in Indo-Iranian.

5.2 *Iranian*

In contrast to the situation in Indo-Aryan, Iranian languages appear to have developed classifiers more frequently; classifiers emerge on nearly a dozen independent branches within Iranian. The majority of these developments are preceded by extended stages with optional plural marking; ultimately, the state of affairs characterized by the GSS generalization appears to have been more active in Iranian than in Indo-Aryan. Our results are also fully in line with the historical record. For example, the development of full-fledged numeral classifiers in Persian was preceded by a prolonged period of optional plural marking. Our evolutionary model suggests that this was the case in several additional lineages. While these findings support the GSS generalization, contact is likely to have played a role as well, especially since Iranian languages are spoken near Turkic languages with classifiers.

Turkic and Iranian languages came increasingly into contact in post-Sasanian times. However, numeral classifiers are likely to be a relatively late development, post-dating the onset of the Turkic expansion from about the 5th century CE (Yunusbayev et al. 2015). The Old Turkic (ca. 7th to 11th century) corpus contains no sortal numeral classifiers (Erdal 2004:226), while the later literary language Chagatay (ca. 14th to 19th century) in Central Asia exhibits certain classifiers, such as *baš* 'head' (Bodrogligeti 2001:155), and classifier use in modern Turkic languages is widespread, found among different branches of Turkic, including Kipchak (e.g., Tatar, Chen Zongzhen & Yi Liqian 1986:70), Oghuz (e.g., Turkmen, Clark 1998:169), and Karluk (e.g., Uzbek, Beckwith 1998).

It is not clear at present whether this distribution reflects a widespread late diffusion throughout Turkic, the effects of some sort of drift or slant-like tendency, or the inheritance of a chronologically deep feature. Complicating matters, Turkic classifiers are often Iranian loans; only in restricted occurrences do Iranian languages borrow Turkic classifiers (e.g., Sariqoli ← Uyghur). Further research is needed to address the question of whether Turkic classifiers are due to Iranian influence, Iranian classifiers are due to Turkic influence, or both groups developed classifiers as a response to loosening restrictions on plural marking under the influence of widespread multilingualism. Importantly, all of these competing scenarios attribute an important role to Turkic/Iranian contact history.

In some cases, the presence of numeral classifiers in Iranian languages points to influence from other Iranian languages. The numeral classifiers found in Wakhi, Yaghnobi and certain Pamir languages are identifiable as Tajik. Tajik's importance as a lingua franca in the region makes borrowing into Iranian particularly plausible. Pashto has likely borrowed the classifier *tana* (m.)/*teni* (f.) from a Turkic language, though it has been in the language long enough to be integrated into the gender system and participate in certain phonological processes.

In other cases, the role of contact is less clear. Kumzari, located in Oman and separated from other Iranian languages by the Persian Gulf, exhibits the classifiers *-ta* and *-kas* (for human beings), formally identical to the Persian classifier *tā* and the Persian word *kas* 'person'. Kumzari is phylogenetically very close to Old, Middle, and modern Persian (Skjærvø 1989), though it is not clear that it is a descendant of Old or Middle Persian. Given the relative isolation of Kumzari, it is possible that it developed these classifiers in parallel with Persian due to parallel drift, but contact with another Iranian language is also a possibility, since there has been longstanding migration to Oman from the other side of the Persian Gulf (Barth 1983).[15]

The remainder of Iranian languages with classifiers (e.g., Mazandarani, Taleshi, and Gilaki) share a core group of classifiers that are formally near-identical to Persian ones (e.g., *tā*), but as in the case of Kumzari, it is difficult to determine on the basis of sound change whether these forms are inherited or borrowed; at the same time, the dominance of Persian over these languages is

15 Although the historical phonology of Kumzari is poorly understood, there is some evidence that it preserves the consonant of Old Iranian final *-*aka*-, given the "etymologically latent *k*" in the definite form *martk-ō* (van der Wal Anonby 2015:38) < **marta-ka*-. Modern Persian does not preserve this consonant (cf. *tā* < Middle Persian *tāg* 'piece', most likely going back to a form **tāka*-). However, Kumzari seems to show preservation only in contexts where the form is suffixed, making it impossible to determine whether the loss of *-k* in *-ta* is regular or reflects a Persian borrowing.

well established (Borjian 2009), lending circumstantial evidence to the notion that they are Persian loans. Some Northwest Iranian languages have classifiers not found in Persian, e.g., Taleshi *gəla* (Paul 2011, Stilo 2018), perhaps cognate with Judeo-Tati *gile* 'time, instance' (Authier 2012:310).

6 Conclusion

Our findings suggest that the emergence of classifiers in Indo-Iranian is tied to optional plural marking, partially explainable by a statistical universal principle in line with the Greenberg-Sanches-Slobin generalization. At the same time, we also find that Indo-Iranian classifiers emerged under contact, reflecting local history and the contingencies of migration and trade. These results seem contradictory when one approaches the distribution of linguistic structures from the popular view that conceptualizes universal pressure and areal histories as conflicting, confounding, and competing factors. The contradiction is resolved, however, if we adopt the view from Distributional Typology (Bickel 2015) where emphasis is placed on the *interaction* between universal and areal factors in shaping synchronic distributions (Bickel 2017). Interestingly, a closer inspection sheds light on different patterns of development across the closely related Indo-Aryan and Iranian branches, reinforcing the need for detailed examination of variation in development across lineages in phylogenetic linguistic work which may be accounted for via different functional and event-based theories.

From a distributional perspective, the borrowing of classifiers (as matter or pattern) is not the mere product of historical contingency. Instead, this borrowing is driven by the decay in number marking, i.e., it follows the Greenberg-Sanches-Slobin generalization. Such a scenario explains our finding that classifiers were considerably more likely to be borrowed in the absence than in the presence of number marking. Areal effects alone cannot explain this difference, while universal effects alone cannot explain why classifiers entered through borrowing rather through spontaneous developments (e.g. by reanalyzing nominal juxtapositions, as in Hackstein's 2010 theory).

While our study supports a scenario of an interaction between universal and areal effects, we caution that our simplified coding scheme may not have picked up all relevant factors and that further research is needed to consolidate our conclusions. What is arguably the most urgent extension is a more fine-grained coding that captures the distribution of number marking over specific noun types, controlling for potential effects of a referential scale.

Acknowledgments

This research was funded by Swiss National Science Foundation (SNSF) Grant Nr. 100015_170241 and the NCCR Evolving Language, SNSF Agreement Nr. 51NF40_180888. We thank audiences at the 13th Meeting of the Association for Linguistic Typology in Pavia for helpful feedback.

References

Acharya, Jayaraj. 1991. *A Descriptive Grammar of Nepali and an Analyzed Corpus*. Washington D.C.: Georgetown University Press.

Acquaviva, Paolo. 2004. The morphosemantics of transnumeral nouns. In Geert Booij, Emiliano Guevara, Angela Ralli, Salvatore Sgroi & Sergio Scalise (eds.), *Morphology and linguistic typology. On-line proceedings of the fourth mediterranean morphology meeting* (*MMM4*) http://mmm.lingue.unibo.it/mmm-proc/MMM4.

Aikhenvald, Alexandra Y. 2000. *Classifiers. A Typology of Noun Categorization Devices* Oxford Studies in Typology and Linguistic Theory. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. & R.M.W. Dixon. 1998. Dependencies between grammatical systems. *Language* 74. 56–80.

Almeida, Matthew. 1989. *A description of Konkani*. Panaji: Thomas Stevens Konkani Kendra.

Anonby, Erik & Ashraf Asadi. 2014. *Bakhtiari Studies: Phonology, Text, Lexicon*. Uppsala: Acta Universitatis Upsaliensis.

Authier, G. 2012. *Grammaire juhuri, ou judeo-tat, langue iranienne des juifs du caucase de l'est*. Wiesbaden: Dr. Ludwig Reichert Verlag.

Azami, Cheragh Ali & Gernot Windfuhr. 1972. *A Dictionary of Sangesari with a Grammatical Outline*. Tehran: Franklin Book Programs.

Baart, Joan L.G. & Muhammad Zaman Sagar. 2004. *Kalam Kohistani Texts*. Islamabad: National Institute of Pakistan Studies, Quaid-i-Azam University.

Barker, Muhammad Abd-al-Rahman. 1969. *A Course in Baluchi*. Montreal: Institute of Islamic Studies, McGill University.

Barth, Fredrik. 1983. *Sohar: Culture and Society in an Omani Town*. Baltimore: Johns Hopkins University Press.

Barz, R.K. & A.V.N. Diller. 1985. Classifiers and standardisation: Some South and South-East Asian comparisons. *Papers in South-East Asian Linguistics* 9. 155–184.

Bauer, Brigitte. 2017. *Nominal Apposition in Indo-European*. Berlin, Boston: De Gruyter.

Beckwith, Christopher I. 1998. Noun specification and classification in Uzbek. *Anthropological Linguistics* 40 (1). 124–140.

Bhatia, Tej K. 1993. *Punjabi: A Cognitive-Descriptive Grammar*. London: Routledge.

Bickel, Balthasar. 2003. Belhare. In Graham Thurgood & Randy J. LaPolla (eds.), *The Sino-Tibetan languages*, 546–570. London: Routledge.

Bickel, Balthasar. 2015. Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis, 2nd edition*, 901–923. Oxford: Oxford University Press.

Bickel, Balthasar. 2017. Areas and universals. In Raymond Hickey (ed.), *The Cambridge Handbook of Areal Linguistics*, 40–55. Cambridge: Cambridge University Press.

Bisang, Walter. 2002. Classification and the evolution of grammatical structures: A universal perspective. *STUF—Language Typology and Universals* 55. 289–308.

Bisang, Walter. 2017. Classification between grammar and culture: A cross-linguistic perspective. In Tanja Pommerening & Walter Bisang (eds.), *Classification from Antiquity to Modern Times*, Berlin: De Gruyter.

Blau, Joyce. 1980. *Manuel de kurde* (*dialecte sorani*). Paris: Librairie C. Klincksieck.

Bodrogligeti, András J.E. 2001. *A Grammar of Chagatay*. Munich: Lincom Europa.

Bollback, Jonathan P. 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7. 88.

Borah, Gautam. 2012. Classifiers in Assamese: Their grammar and meaning chains. In Gwendolyn Hyslop, Stephen Morey & Mark W. Post (eds.), *North East Indian Linguistics* vol. 4, 293–314. New Delhi: Foundation Books.

Borer, Hagit. 2005. *Structuring Sense, Volume I: In Name Only*. Oxford: Oxford University Press.

Borjian, Habib. 2009. Median succumbs to Persian after three millennia of coexistence: Language shift in the Central Iranian Plateau. *Journal of Persianate Studies* 2. 62–87.

Burghart, Richard. 1992. *Introduction to Spoken Maithili in the Social Context*. Heidelberg: Südasien Institut, Abteilung für Ethnologie. 3 vols.

Cardona, George. 1965. *A Gujarati Reference Grammar*. Philadelphia: University of Pennsylvania Press.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76 (1).

Cassirer, Ernst. 1923. *Philosophie der symbolischen Formen. Erster Teil: Die Sprache*. Berlin: Bruno Kassierer Verlag.

Cathcart, Chundra. 2015. *Iranian Dialectology and Dialectometry*: University of California, Berkeley dissertation.

Chandralal, Dileep. 2010. *Sinhala*. Amsterdam, Philadelphia: John Benjamins.

Chatterji, Suniti Kumar. 1926. *The Origin and Development of the Bengali Language*. Calcutta: Calcutta University Press.

Chen Zongzhen & Yi Liqian. 1986. *Tata'er jianzhi*. Peking: Minzu chubanshe.

Chowdhary, Runima. 2012. On classifiers in Asamiya. In Gwendolyn Hyslop, Stephen Morey & Mark W. Post (eds.), *North East Indian Linguistics* vol. 4, 267–291. New Delhi: Foundation Books.

Clark, Larry. 1998. *Turkmen Reference Grammar*. Wiesbaden: Harrassowitz.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Cristofaro, Sonia. 2012. Cognitive explanations, distributional evidence, and diachrony. *Studies in Language* 36. 645–670.

Croft, William. 1994. Semantic universals in classifier systems. *Word* 45 (2). 145–171.

David, Anne Boyle. 2015. *Descriptive Grammar of Bangla*. Berlin, Boston: De Gruyter Mouton.

Deo, Ashwini. 2018. Dialects in the Indo-Aryan landscape. In Charles Boberg, John Nerbonne & Dominic Watt (eds.), *The Handbook of Dialectology*, 535–546. Oxford: John Wiley & Sons.

Dhongde, Ramesh Vaman & Kashi Wali. 2009. *Marathi*. Amsterdam, Philadelphia: John Benjamins.

Doetjes, Jenny. 2012. Count/mass distinctions across languages. In C. Maienborn, V. von Heusinger & Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning*. *Volume 3*, 2559–2580. Berlin: De Gruyter.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473 (7345). 79–82.

Durkin-Meisterernst, Desmond. 2009. Khwarezmian. In Gernot Windfuhr (ed.), *The Iranian Languages*, chap. 6, 336–376. London: Routledge.

Durkin-Meisterernst, Desmond. 2014. *Grammatik des Westmitteliranischen* (*Parthisch und Mittelpersisch*). Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Egorova, R.P. 1966. *Jazyk sindxi*. Moscow: Akademia Nauk SSSR.

Emeneau, Murray B. 1956. India as a linguistic area. *Language* 32. 3–16.

Emeneau, Murray B. 1965 [1980]. India and linguistic areas. In Anwar S. Dil (ed.), *Language and Linguistic Area: Essays by Murray B. Emeneau*, 126–166. Stanford, CA: Stanford University Press.

Emmerick, Ronald. 1989. Khotanese and Tumshuqese. In Rüdiger Schmitt (ed.), *Compendium Linguarum Iranicarum*, 204–229. Wiesbaden: Dr. Ludwig Reichert Verlag.

Endresen, Rolf Theil & Knut Kristiansen. 1981. Khowar studies. *Acta Iranica* 21. 210–243.

Erdal, Marcel. 2004. *A Grammar of Old Turkic*. Leiden: Brill.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Gebhardt, Lewis. 2018. Accounting for **yek ta* in Persian. In Alireza Korangy & Corey Miller (eds.), *Trends in Iranian and Persian Linguistics*, 213–232. Berlin/Boston: Mouton de Gruyter.

Geiger, Wilhelm. 1942. *Beiträge zur singhalesichen Sprachgeschichte*. Munich: Verlag der Bayerischen Akademie der Wissenschaften.

Gelman, Andrew & Cosma Rohilla Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66. 8–38.

Gil, David. 1987. Definiteness, noun-phrase configurationality, and the count-mass distinction. In E.J. Reulen & A.G.B. ter Meulen (eds.), *The Representation of (In)definiteness*, 254–269. Cambridge: MIT Press.

Gil, David. 2013. Numeral classifiers. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/55.

Gnandesikan, Amalia E. 2017. *Dhivehi: The Language of the Maldives*. Berlin, Boston: De Gruyter Mouton.

Golden, Peter. 2006. Turks and Iranians: an historical sketch. In Lars Johanson & Christiane Bulut (eds.), *Turkic-Iranian Contact Areas: Historical and Linguistic Aspects*, 17–38. Wiesbaden: Harrassowitz.

Greenberg, Joseph. 1972. Numeral classifiers and substantival number: Problems in the genesis type. *Working Papers on Language Universals* 9. 1–39.

Grierson, George A. 1929. *Torwali: An Account of a Dardic Language of the Swat Kohistan*. London: Royal Asiatic Society.

Grinevald, Colette. 2000. A morphosyntactic typology of classifiers. In Gunter Senft (ed.), *Systems of Nominal Classification*, 50–92. New York: Cambridge University Press.

Grjunberg, Aleksandr Leonovič & Ivan Michajlovič Steblin-Kamenskij. 1988. *La langue wakhi*. Paris: Editions de la Maison des Sciences de l'Homme.

Gusain, Lakhan. 2003. *Mewati*. München: Lincom Europa.

Gusain, Lakhan. 2004. *Marwari*. München: Lincom Europa.

Hackstein, Olav. 2010. *Apposition and Nominal Classification in Indo-European and Beyond*. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

Haspelmath, Martin. 2013. Occurrence of nominal plurality. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/34.

Haspelmath, Martin. 2018. Toward a new conceptual framework for comparing gender systems and some so-called classifier systems. Talk presented at Stockholm University, Department of Linguistics on April 13, 2018.

Heston, Wilma Louise. 1980. Some areal features: Indian or Irano-Indian. *International Journal of Dravidian Linguistics* 9 (1). 141–157.

Hoffmann, Karl & Bernhard Forssman. 2004. *Avestische Laut- und Flexionslehre* vol. 84 Innsbrucker Beiträge zur Sprachwissenschaft. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck 2nd edn.

Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck & Fredrik Ronquist. 2016. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* 65 (4). 726–736.

Huelsenbeck, John P., Rasmus Nielsen & Jonathan P. Bollback. 2003. Stochastic mapping of morphological characters. *Systematic Biology* 52 (2). 131–158.

Igla, Birgit. 1996. *Das Romani von Ajia Varvara: deskriptive und historisch-vergleichende Darstellung eines Zigeunerdialekts*. Wiesbaden: Harrassowitz.

Imai, M. & D. Gentner. 1997. A crosslinguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition* 62. 169–200.

Inglis, Douglas. 2007. *Nominal Structure in Tai Khamti* vol. 312 Research Paper. Payap University.

Ioannesjan, Ju. A. 1999. *Geratskij dialekt jazyka dari*. Moscow: Izdatel'skaja firma "vostočnaja literatura" RAN.

Jäger, Gerhard. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Reports* 5. https://www.nature.com/articles/sdata2018189.

Katenina, T.E. 1963. *Jazyk maratxi*. Moscow: Izdatel'stvo vostočnoj literatury.

Kent, Roland. 1951. *Old Persian* vol. 33 American Oriental Series. New Haven: American Oriental Society.

Kieffer, Charles. 2003. *Grammaire de l'ōrmuṛī de baraki-barak* (*lōgar, afghanistan*). Wiesbaden: Dr. Ludwig Reichert Verlag.

Kieffer, Charles M. 2009. Parachi. In Gernot Windfuhr (ed.), *The Iranian Languages*, 693–720. London, New York: Routledge.

Kiseleva, L.N. 1985. *Jazyk dari afganistana*. Moscow: Izdatel'stvo Nauka (glavnaja redakcija vostočnoj literatury).

Krifka, Manfred. 1995. Common nouns: A contrastive analysis of Chinese and English. In Gregory Carlson & Francis Jeffery Pelletier (eds.), *The Generic Book*, 398–411. Chicago: University of Chicago Press.

Krishnamurti, Bhadriraju & J.P.L. Gwynn. 1985. *A Grammar of Modern Telugu*. Oxford: Oxford University Press.

Kruschke, John K. 2011. Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science* 6. 299–312.

Kuz'mina, Elena. 2007. *The Origin of the Indo-Iranians*. Leiden: Brill.

Lambert, H.M. 1943. *Marathi Language Course*. London: Oxford University Press.

Lazard, Gilbert. 1963. *La langue des plus anciens monuments de la prose persane*. Paris: Klincksieck.

Lehmann, Christian. 2000. On the German numeral classifier system. In Chris Schaner-Wolles, John R. Rennison & Friedrich Neubarth (eds.), *Naturally! Linguistic Studies in Honour of Wolfgang Ulrich Dressler Presented on the Occasion of his 60th Birthday*, 249–253. Torino: Rosenberg & Sellier.

Lehmann, Christian. 2010. On the function of numeral classifiers. In Franck Floricic (ed.), *Essais de typologie et de linguistique générale. Mélanges offerts à Denis Creissels*, 435–445. Lyon: École Normale Supérieure.

Lehmann, Thomas. 1993. *A Grammar of Modern Tamil* vol. 1 PILC Publication. Pondicherry Institute of Linguistics and Culture, India: Pondicherry, India: Pondicherry Institute of Linguistics and Culture 2nd edn. Includes index Includes bibliographical references (p. [379]–381).

Liljegren, Henrik. 2016. *A Grammar of Palula*. Berlin: Language Science Press.

Lorimer, David Lockhart Robertson. 1939. *The Ḍumāki language: Outlines of the Speech of the Ḍoma, or Bērīcho, of Hunza*. Nijmegen: Dekker and van de Vegt.

Lucy, John A. 1992. *Grammatical Categories and Cognition. A Case Study of the Linguistic Relativity Hypothesis*. Cambridge: Cambridge University Press.

Macdonell, Arthur Anthony. 1910. *Vedic Grammar*. Strassburg: K.J. Trübner.

Mache, Avazeh. 2012. *Numeral Classifiers in Persian*. München: Lincom Europa.

MacKenzie, David Neil. 1971. *A Concise Pahlavi Dictionary*. London: Oxford University Press.

MacKinnon, Colin. 2003. The dialect of Xorramābād and comparative notes on other Lor dialects. *Studia Iranica* 31. 103–138.

Maddison, Wayne P. & Richard G. Fitzjohn. 2014. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology* 64. 127–136.

Mahootian, Shahrzad. 1997. *Persian*. London: Routledge.

Masica, Colin P. 1991. *The Indo-Aryan Languages*. Cambridge: Cambridge University Press.

Matisoff, James. 1978. *Variational Semantics in Tibeto-Burman: The 'Organic' Approach to Linguistic Comparison* vol. 6 Occasional Papers of the Wolfenden Society on Tibeto-Burman Linguistics. Philadelphia: Institute for the Study of Human Issues.

Matras, Yaron. 2002. *Romani—A Linguistic Introduction*. Cambridge: Cambridge University Press.

Matras, Yaron. 2012. *A Grammar of Domari*. Berlin: De Gruyter Mouton.

Matras, Yaron & Jeannette Sakel. 2007. Investigating the mechanisms of pattern replication in language convergence. *Studies in Language* 31 (4). 829–865.

McElreath, Richard. 2020. *Statistical rethinking: A Bayesian Course with Examples in R and Stan*. CRC press.

Metzger, Mathias. 2003. *Die Sprache der Vakīl-Briefe aus Rājasthān*. Heidelberg: Ergon.

Moazami, Mahnaz. 2014. *Wrestling with the Demons of the Pahlavi Widēwdād*. Leiden & Boston: Brill.

Morgenstierne, Georg. 1929. *Indo-iranian frontier languages: Vol. 1, parachi and ormuri*. Oslo: H. Aschehoug & Co.

Mukherjee, Tarapada. 1963. *The Old Bengali Language and Text*. Calcutta: University of Calcutta.

Nawata, Tetsuo. 1984. *Mazandarani* vol. 17 Asian and African Grammatical Manual. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Neukom, Lukas & Manideepa Patnaik. 2003. *A Grammar of Oriya* (ASAS 17). Universität Zürich.

Nielsen, Rasmus. 2002. Mapping mutations on phylogenies. *Systematic Biology* 51 (5). 729–739.

Nitz, Eike & Sebastian Nordhoff. 2010. Subtractive plural morphology in Sinhala. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rara & Rarissima*, 247–266. Berlin: De Gruyter.

Oberlies, Thomas. 2001. *Pali: A Grammar of the Language of the Theravada Tipitaka. With a Concordance to Pischel's Grammatik der Prakrit-Sprachen*. Berlin: Walter de Gruyter.

Oberlies, Thomas. 2005. *A historical grammar of hindi*. Graz: Leykam.

Oldenberg, Hermann. 1909. *gveda: Textkritische und exegetische Noten, erstes bis sechstes Buch*. Berlin: Weidmannsche Buchhandlung. Rpr. 1970. Nendeln, Liechtenstein: Kraus.

Pagel, Mark. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London B* 255. 37–45.

Pagel, Mark & Andrew Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by Reversible-Jump Markov Chain Monte Carlo. *The American Naturalist* 167 (6). 808–825.

Pandharipande, Rajeshwari. 1997. *Marathi*. London: Routledge.

Paul, Daniel. 2011. *A Comparative Dialectal Description of Iranian Taleshi*: University of Manchester dissertation.

Paul, Ludwig. 1998. *Zazaki: Grammatik und versuch einer dialektologie* vol. 18 Beiträge zur Iranistik. Wiesbaden: Dr. Ludwig Reichert Verlag.

Paxalina, T.N. 1959. *Iškašimskij jazyk*. Moscow: Nauka.

Paxalina, T.N. 1971. *Sarykol'sko-russkij slovar'*. Moscow: Nauka.

Penzl, Herbert. 1955. *A grammar of Pashto: A dDescriptive Study of the Dialect of Kandahar, Afghanistan*. Washington D.C.: American Council of Learned Societies.

Petersen, Jan Heegård. 2015. Kalasha texts—with introductory grammar. *Acta Linguistica Hafniensia* 47. 1–275.

Quine, W.V.O. 1960. *Word and Object*. Cambridge, MA: Technology Press of the Massachusetts Institute of Technology.

Rastorgueva, V.S., A.A. Kerimova, A.K. Mamedzade, L.A. Pireiko & Joy I. Edelman. 2012. *The Gilaki Language*. Uppsala: Uppsala Universitet.

Revell, Liam J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3. 217–223.

Saksena, Baburam. 1971. *Evolution of Awadhi* (*a Branch of Hindi*). Delhi: Motilal Banarsidass.

Sanches, Mary & Linda Slobin. 1973. Numeral classifiers and plural marking: An implicational universal. *Working Papers on Language Universals* 11. 1–23.

Schiffmann, Harold F. 1999. *A Reference Grammar of Spoken Tamil*. Cambridge: Cambridge University Press.

Schmidt, Ruth Laila, Razwal Kohistani & Mohammad Manzar Zarin. 2008. *A Grammar of the Shina Language of Indus Kohistan*. Wiesbaden: Harrassowitz.

Schmidtke-Bode, Karsten. 2019. Introduction. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the Evidence*, iii–xii. Berlin: Language Science Press.

Shackle, Christopher. 1976. *The Siraiki Language of Central Pakistan: A Reference Grammar*. London: School of Oriental and African studies, University of London.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In R.M.W. Dixon (ed.), *Grammatical Relations in Australian Languages*, 112–171. Canberra: Australian Institute of Aboriginal Studies.

Simpson, Andrew, Hooi Ling Soh & Hiroki Nomoto. 2011. Bare classifiers and definiteness: A cross-linguistic investigation. *Studies in Language* 35 (1). 168–193.

Sims-Williams, Nicholas. 2007. *Bactrian Documents from Northern Afghanistan, Vol. 2: Letters and Buddhist Texts*. London: The Nour Foundation in Association with Azimuth Editions.

Skjærvø, Prods Oktor. 1989. Languages of Southeast Iran. In Rüdiger Schmitt (ed.), *Compendium Linguarum Iranicarum*, 363–369. Dr. Ludwig Reichert Verlag.

Skjærvø, Prods Oktor. 2009. Middle West Iranian. In Gernot Windfuhr (ed.), *The Iranian languages*, 196–278. London: Routledge.

Southworth, Franklin. 1962. Review of A Grammar of Old Marathi by Alfred Master. *Language* 46. 507–513.

Stilo, Don. 2018. Numeral classifier systems in the praxes-iran linguistic area. In William B. McGregor & Søren Wichmann (eds.), *The Diachrony of Classification Systems*, 135–164. Amsterdam: Benjamins.

Tang, Marc & One-Soon Her. 2019. Insights on the Greenberg-Sanches-Slobin generalization: Quantitative typological data on classifiers and plural markers. *Folia Linguistica* 53 (2). 297–331.

Tavaré, Simon. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17 (2). 57–86.

Thomas, Bertram. 1930. *The Kumzari Dialect of the Shihuh Tribe, Arabia and a Vocabulary*. London: The Royal Asiatic society.

Thomason, Sarah Grey & Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley & Los Angeles: University of California Press.

Thordarson, Fridrik. 2008 [2009]. *Ossetic Grammatical Studies* (Veröffentlichungen zur Iranistik 48). Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

Tirkey, Masato Kobayashi & Bablu. 2017. *The Kurux Language: Grammar, Texts, and Lexicon* vol. 08 Brill's Studies in South and Southwest Asian Languages. Leiden: Brill.

Tiwari, Udai Narain. 1960. *The Origin and Development of Bhojpuri*. Kolkata: The Asiatic Society.

Tulpule, S.C. 1963. *Prācīn marāṭhī korīv lekh* (*Old Marathi inscriptions*). Poona: Poona University Press.

Turner, Ralph L. 1962–1966. *A Comparative Dictionary of Indo-Aryan Languages*. London: Oxford University Press.

Ucida, Norihoko. 1979. *Oral Literature of the Saurashtrans*. Calcutta: Simant.

van der Wal Anonby, Christina. 2015. *A Grammar of Kumzari: A Mixed Perso-Arabian Language of Oman*: Rijksuniversiteit te Leiden dissertation.

Wali, Kashi & Omkar N. Koul. 1996. *Kashmiri*. London, New York: Routledge.

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2018. The ASJP database (version 18). http://asjp.clld.org/.

Windfuhr, Gernot & John J. Perry. 2009. Persian and Tajik. In Gernot Windfuhr (ed.), *The Iranian Languages*, 416–544. London: Routledge.

Woolner, Alfred C. 1928. *Introduction to Prakrit*. Delhi: Motilal Banarsidass.

Xromov, Al'bert L. 1972. *Jagnobskij jazyk*. Moscow: Nauka.

Yar-shater, Ehsan. 1969. *A Grammar of Southern Tati Dialects* (Median Dialect Studies 1). The Hague: Mouton.

Yoshida, Yutaka. 2009. Sogdian. In Gernot Windfuhr (ed.), *The iranian languages*, 279–335. London, New York: Routledge.

Yunusbayev, Bayazit, Mait Metspalu, Ene Metspalu, Albert Valeev, Sergei Litvinov, Ruslan Valiev, Vita Akhmetova, Elena Balanovska, Oleg Balanovsky, Shahlo Turdikulova, Dilbar Dalimova, Pagbajabyn Nymadawa, Ardeshir Bahmanimehr, Hovhannes Sahakyan, Kristiina Tambets, Sardana A. Fedorova, Nikolay Barashkov, Irina Khidiyatova, Evelin Mihailov, Rita Khusainova, Larisa Damba, Miroslava Derenko, Boris Malyarchuk, Ludmila P. Osipova, Mikhail I. Voevoda, Levon Yepiskoposyan, Toomas Kivisild, Elza Khusnutdinova & Richard Villems. 2015. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genetics* 11 (4). 1–24.

Zarubin, I.I. 1960. *Šugnanskije teksty i slovar'*. Moscow: Izdatel'stvo Akademij Nauk SSSR.

Appendix: Language states

This section contains codings for languages used in this study. We make the assumption that if an exhaustive search of a grammar does not turn up any reference to unambiguous sortal numeral classifiers, then they are absent from the language in question.

A.1 *Agia Varvara Romani* [*vlax1238*] (*+clf,+obl.pl*)

Singular forms are always morphologically distinct from their plural counterparts, usually via the addition of a plural suffix or alternation of the final vowel (Igla 1996:23 ff.). When a numeral greater than one modifies a noun, the noun is morphologically plural (p. 45). When the counted item consists of indefinite objects, then *-tane* (< Turkish *tane* 'piece, part') appears next to the numeral; anaphoric use of this classifier is obligatory (p. 45).

A.2 *Assamese* [*assa1263*] (*+clf,-obl.pl*)

Assamese has a large inventory of sortal classifiers. Plural marking is optional (Borah 2012, Chowdhary 2012).

A.3 *Avestan* [*aves1237*] (*-clf,+obl.pl*)

Plural number is consistently marked, given rich agreement morphology (Hoffmann & Forssman 2004).

A.4 *Awadhi* [*awad1243*] (*+clf,-obl.pl*)

Classifiers are present; information regarding plural marking is difficult to extract from Saksena 1971:115 ff., but it appears to be optional.

A.5 *Bactrian* [*bact1239*] (*-clf,-obl.pl*)

In late Bactrian, case and number distinctions have been neutralized due to the loss of distinctions between final vowels, resulting in "an unmarked form without ending … which may be used with either sg. or pl. reference, and a marked pl. form" (Sims-Williams 2007:40).

A.6 *Bagri* [*bagr1243*] (*-clf,morph.pl*)

There are at least three declensional classes, in one of which the plural and singular direct forms are identical. The distinction between animacy and inanimacy, rather than ending in *-i* versus other segments, is made by the author, who explicitly states that the suffix *-ã* is optional on animate nouns. This is almost akin to a mixture of a system like that of Hindi, where plural cannot be marked on some noun case forms, and a system where plural marking is truly optional.

A.7 *Bakhtiyari* [*bakh1245*] (*+clf,-obl.pl*)
Classifiers are present, and plural marking is variable (Anonby & Asadi 2014).

A.8 *Bengali* [*beng1280*] (*+clf,-obl.pl*)
Bengali has a large repertoire of numeral classifiers (David 2015:135). Classifiers are obligatory with non-numeric quantifiers and lower numbers; optional with numbers ending in 'hundred', 'thousand', 'lakh', etc. (p. 142). Numeral classifiers cannot cooccur with nouns denoting a countable unit, e.g., units of weight, currency, time, except in certain emphatic contexts (p. 142). Plural marking is non-obligatory (p. 76).

A.9 *Bhojpuri* [*bhoj1244*] (*+clf,-obl.pl*)
Numeral classifiers are present, and plural marking is non-obligatory (Tiwari 1960:120, 228, 230)

A.10 *Dari* [*dari1249*] (*+clf,-obl.pl*)
Classifier use is common, but not obligatory; plural marking is optional; it is not clear if plural marking can co-occur with classifier use as in Standard Modern Persian (Kiseleva 1985:74–75; Ioannesjan 1999:58–59).

A.11 *Dhivehi* [*dhiv1236*] (*-clf,-obl.pl*)
For nonhuman nouns, plural marking is optional when plurality is clear from context (Gnandesikan 2017:59).

A.12 *Domari* [*doma1258*] (*-clf,-obl.pl*)
Plural marking on enumerated nouns interacts significantly with whether the numbers are inherited Indo-Aryan forms or borrowed from Arabic (see Matras 2012:97, 188 ff.). Plural number is optionally marked on nouns modified by 2–3 (inherited numbers), obligatory on nouns modified by 4–10 (Arabic numbers), and optionally marked on nouns from 11 upward (Arabic numbers).

A.13 *Dumaki* [*doma1260*] (*-clf,+obl.pl*)
For virtually all nouns, the singular form is distinguishable from the plural form (Lorimer 1939:24 ff.); this is achieved via suffixation or a stem alternation.

A.14 *Gilaki* [*gila1241*] (*+clf,-obl.pl*)
Rastorgueva et al. (2012) list several classifiers. Classifier use is optional when enumerating nouns, but appears to be quite common and obligatory in anaphoric use. In all examples given of counted nouns, there is no overt plural

marking on the head noun (regardless of whether a classifier is present). Plural can otherwise be marked by means of certain suffixes.

A.15 *Gujarati* [*guja1252*] (*-clf,-obl.pl*)
Cardona (1965:66–67) refers to the plural marker *-o* as "optional," and it seems to largely be omitted when nouns are modified by a number greater than 1; so-called "variable" nouns display a special "dependent stem form" when they are semantically plural, regardless of the presence of the suffix *-o*.

A.16 *Hindi* [*hind1269*] (*-clf,morph.pl*)
Hindi shows four declensional classes; the details of number marking are different for each one. Plural marking is not morphologically possible on C-final masculine nouns in direct case (Oberlies 2005).

A.17 *Iron Ossetic* [*osse1243*] (*-clf,+obl.pl*)
Plural number is marked on nouns by means of the suffix *-t-* (Thordarson 2008 [2009]:117). In most contexts (except for contexts of enumeration, see below), use of *-t-* appears to be obligatory. When a noun is enumerated by a numeral greater than one, the noun is marked by the suffix *-i* (Digor), identical to the genitive suffix. According to Thordarson (2008 [2009]:132), this suffix continues the Old Iranian plural suffix **-ah*. Nouns enumerated by numbers greater than one are always marked in a way that renders them distinct from singular nouns.

A.18 *Ishkashimi* [*ishk1246*] (*+clf,-obl.pl*)
Ishkashimi contains at least three classifiers; nouns modified by a numeral greater than one can appear in singular or plural form (Paxalina 1959:50).

A.19 *Judeo-Tati* [*jude1256*] (*-clf,-obl.pl*)
Plural is marked on nouns with the suffix *-ho* (Authier 2012:79). Overt plural marking on enumerated nouns seems virtually non-existent.

A.20 *Kalam Kohistani* [*indu1241*] (*-clf,+obl.pl*)
Kalam Kohistani achieves plural marking on a number of nouns via a vowel fronting process, which also is found in oblique forms of nouns, and appears to mark plural consistently on nouns (Baart & Sagar 2004:21). The word *khur* 'foot' may show variability in plural marking, but it is not clear from the data given.

A.21 *Kalasha* [*kala1372*] (*-clf,-obl.pl*)
Plural marking is optional (Petersen 2015:35–36).

A.22 *Kashmiri* [*kash1277*] (*-clf,morph.pl*)
According to Wali & Koul (1996:190 ff.): "plurals are formed from singular stems by vowel change, palatalization and suffixation. A few nouns stay invariant. Masculine plurals are formed differently than the feminine plurals." Mass nouns, most body parts, and borrowed English nouns use the same forms in both the singular and the plural. Masculine nouns do not change for plurality if they have certain phonotactic properties or are borrowed from Hindi/Urdu and English with a final consonant.

A.23 *Khotanese Saka* [*khot1251*] (*-clf,+obl.pl*)
Plural number is consistently marked, given rich agreement morphology (Emmerick 1989).

A.24 *Khowar* [*khow1242*] (*-clf,-obl.pl*)
Plural marking appears to be optional on the basis of examples provided in Endresen & Kristiansen (1981).

A.25 *Khwarezmian* [*khwa1238*] (*-clf,+obl.pl*)
Plural is consistently marked (Durkin-Meisterernst 2009).

A.26 *Konkani* [*konk1267*] (*-clf,morph.pl*)
Certain noun categories have identical singular and plural endings in the direct case; otherwise, plural is consistently marked (Almeida 1989:126 ff.)

A.27 *Kumzari* [*kumz1235*] (*+clf,+obl.pl*)
From Thomas's 1930 description, plural marking appears to be obligatory. Kumzari numerals are nearly identical to their Modern Persian cognates; however, from seven upwards, the Kumzari numerals all end in *-tā*, which is analyzed as a suffix. For human beings, a suffix *-kay* attaches to the number one *yek*(*kay*); for two onward, the suffix *-kas* is used. According to a newer description, the numeral classifier *-tā* or *-ta* in Kumzari can also occur on numerals below seven (van der Wal Anonby 2015:47)

A.28 *Luri* [*luri1257*] (*-clf,-obl.pl*)
According to MacKinnon (2003), plural marking is the same as in Modern Persian. No information regarding numeral classifiers is provided.

A.29 *Maithili* [*mait1250*] (*+clf,-obl.pl*)
Maithili has at least two classifiers (Burghart 1992:v. 1, 117). The suffix *-sab*(*h*) is an optional plural marker; when added to nouns that are inherently plural

(e.g., vegetables), takes on the meaning "X and such things." Some other suffixes exist for reference to persons, used in formal speech (Burghart 1992:v. 1, 50–51).

A.30 *Marathi* [*mara1378*] (*-clf,morph.pl*)
Plural number must be marked on semantically plural nouns, except where morphologically impossible, e.g., masculine kinship terms, certain loanwords (Pandharipande 1997:366–367). Emeneau (1956:11) claims that Marathi has a classifier *jaṇ/jaṇī* (f.) that appears "when nouns denoting persons are numerated by numerals higher than four (and optionally for two to four)." Lambert (1943:243) says the following: "When the numerals refer to persons, special forms are used instead of *don*, *tin*, *car*; to other numerals the word *zən* (m. *zən*, cf.fem. *zəna*(*n*); f. *zəni*, cr.fem. *zəni*(*n*)) is usually added. This word is often added also to the special forms of *don*, *tin*, *car*." No examples are given. Katenina (1963:50) gives examples of the special forms *doghe*, *tighe*, *ćaughe*, as well as the forms *ʃaṇ* (m.) and *ʃaṇī* (f.) 'people' which show the latter form as a head noun, but never in close apposition with another (head) noun. The interaction between the numerals and *ʃaṇ*(*ī*) is striking; however, other grammars gloss these special forms simply as 'both', 'the three', and 'the four' respectively (Dhongde & Wali 2009:59). These forms appear in Old Marathi as substantivized numerals, e.g., *he tighe bhāu* 'these three were brothers' (Tulpule 1963, apud Southworth 1962:425).

A.31 *Marwari* [*marw1260*] (*-clf,morph.pl*)
There are at least three declensional classes, in one of which the plural and singular direct forms are identical. Plural number is marked on plural nouns, where morphologically possible (Gusain 2004:20, 29).

A.32 *Mazandarani* [*maza1291*] (*+clf,-obl.pl*)
Classifiers are present, and plural marking is optional (Nawata 1984:9–10).

A.33 *Mewati* [*mewa1249*] (*-clf,morph.pl*)
There are at least three declensional classes, in one of which the plural and singular direct forms are identical. Plural number is marked on plural nouns, where morphologically possible (Gusain 2003:20, 29).

A.34 *Middle Persian* [*pahl1241*] (*-clf,-obl.pl*)
Middle Persian can mark plural with the suffixes *-hā* and *-ān*, but plural is frequently unmarked on plural nouns (Skjærvø 2009:223).

A.35 ***Modern Persian*** **[*midd1350*] (+*clf,-obl.pl*)**
Modern Persian has several numeral classifiers, the most basic and widespread of which is *tā*, optionally used with numbers larger than one (Windfuhr & Perry 2009:478). Plural marking is optional, but the noun being modified can be marked for plural number if it has specific reference (Mahootian 1997:195). Classifiers are obligatory in anaphoric use.

A.36 ***Nepali*** **[*nepa1254*] (+*clf,-obl.pl*)**
Nepali contains several numeral classifiers (Acharya 1991:100); plural number is marked with *-haru*; according to Acharya this marking is optional (pp. 98–99). From Acharya's examples, *-haru* can can co-occur with numeral classifiers (p. 100). According to Bhim Lal Gautam (p.c.), *-haru* is obligatory with human nouns; however, non-human nouns cannot co-occur with overt plural marking and a classifier.

A.37 ***Old East Rajasthani*** **[*dhun1238*] (-*clf,morph.pl*)**
There are at least three declensional classes, in one of which the plural and singular direct forms are identical. Plural number is marked on plural nouns, where morphologically possible (Metzger 2003).

A.38 ***Old Persian*** **[*oldp1254*] (-*clf,+obl.pl*)**
Plural number is consistently marked, given rich agreement morphology (Kent 1951).

A.39 ***Oriya*** **[*oriy1255*] (+*clf,-obl.pl*)**
Oriya has several classifiers; plural marking is optional (Neukom & Patnaik 2003).

A.40 ***Ormuri*** **[*ormu1247*] (+*clf,-obl.pl*)**
According to Kieffer (2003:133), classifiers are used as they are in Dari.

A.41 ***Pali*** **[*pali1273*] (-*clf,+obl.pl*)**
Pali generally maintains a clear morphological distinction between singular and plural; although the nominative singular and plural of ā-stems fell together due to regular sound change, a secondary plural suffix came into use in order to distinguish between the two numbers (Oberlies 2001:150–151)

A.42 ***Palula*** **[*phal1254*] (-*clf,+obl.pl*)**
Plural is consistently marked on count nouns with one of five suffixes, which are accompanied in some cases by stem alternations (Liljegren 2016:103–104).

A.43 ***Panjabi*** **[*panj1256*] (*-clf,morph.pl*)**
There are at least three declensional classes, in one of which the plural and singular direct forms are identical. Plural number is marked on plural nouns, where morphologically possible (Bhatia 1993:214–215).

A.44 ***Parachi*** **[*para1299*] (*-clf,-obl.pl*)**
Little information about interactions between numeral modification and number marking can be found in Kieffer (2009). According to Morgenstierne (1929:50), plural marking is optional, and rare when a numeral modifies the noun. No information regarding classifiers is given.

A.45 ***Parthian*** **[*part1239*] (*-clf,-obl.pl*)**
Durkin-Meisterernst (2014:272) describes a scenario for all of Middle West Iranian whereby plural marking is optional on all enumerated nouns, but animate nouns are more often marked for plural number than inanimates.

A.46 ***Pashto*** **[*pash1269*] (*+clf,+obl.pl*)**
Plural number appears to be consistently marked on Pashto nouns (Penzl 1955:45 ff.); most paradigms are relatively complex and contain stem alternations. Numeral classifiers are possible, and co-occur with nouns marked for plural number, possibly with gender agreement (p. 82).

A.47 ***Prakrit*** **[*maha1305*] (*-clf,+obl.pl*)**
Plural number is consistently marked, given rich agreement morphology (Woolner 1928).

A.48 ***Rakhshani Baluchi*** **[*west2368*] (*-clf,morph.pl*)**
Nouns can be marked for indefiniteness and singularity via the suffix *-e*; otherwise, there is no morphological distinction between singular and plural (Barker 1969:3 ff.).

A.49 ***Sangesari*** **[*sang1315*] (*-clf,morph.pl*)**
Classifiers do not exist in Sangesari. According to Azami & Windfuhr (1972:70 ff.), plural is consistently marked on oblique nouns with the suffix *-uon*, but rarely on direct nouns, except for a restricted set of items.

A.50 ***Sanskrit*** **[*sans1269*] (*-clf,+obl.pl*)**
Plural number is consistently marked, given rich agreement morphology (Macdonell 1910).

A.51 ***Sariqoli*** [***sari1246***] (***+clf,-obl.pl***)
Classifiers are present, and plural marking is optional (Paxalina 1971).

A.52 ***Saurashtran*** [***saur1248***] (***-clf,morph.pl***)
Plural number is marked on Saurashtran nouns by means of two suffixes, *-nu* and *-lu* (< Telugu) (Ucida 1979:45–46). When a numeral greater than one modifies a noun, the noun is plural, but certain non-human nouns (e.g., days, years) do not take plural form.

A.53 ***Shina*** [***shin1264***] (***-clf,morph.pl***)
Plural appears to be consistently marked on count nouns (Schmidt et al. 2008).

A.54 ***Shughni*** [***shug1248***] (***-clf,-obl.pl***)
Plural marking is optional, but classifiers do not appear to be present (Zarubin 1960).

A.55 ***Sindhi*** [***sind1272***] (***-clf,morph.pl***)
There are multiple declensional classes, in one of which the plural and singular direct forms are identical. Irregular plurals can be found for kinship terms. Arabic words often have distinctive plurals borrowed from the source language. Plural number is marked on plural nouns, where morphologically possible (Egorova 1966:27–28).

A.56 ***Sinhala*** [***sinh1246***] (***+clf,+obl.pl***)
When animate nouns are modified by a numeral, the form of the numeral used is different from that which is used when an inanimate noun is modified by a numeral (Chandralal 2010:60). Plural marking is obligatory, when applicable.

A.57 ***Siraiki*** [***sera1259***] (***-clf,morph.pl***)
Certain noun categories have identical singular and plural endings in the direct case; otherwise, plural is consistently marked (Shackle 1976)

A.58 ***Sogdian*** [***sogd1245***] (***-clf,-obl.pl***)
Sogdian heavy stem nouns occasionally show a form that is identical to the singular in plural contexts; this behavior displays variation (Yoshida 2009:313).

A.59 ***Sorani Kurdish*** [***cent1972***] (***-clf,-obl.pl***)
According to Blau (1980:45–46), the simple form of a noun can have either a singular or plural reading. In general, plural number is marked with the suffix *-an*.

A.60 *South Tati* [*esht1238*] (*-clf,-obl.pl*)
According to Yar-shater (1969:78), "nouns modified by a numeral higher than one, or by an expression denoting plurality, are generally expressed in the plural in Chali [a particular dialect]." Occasionally, however, the singular is used. In the other dialects, normally the singular is used for enumerated nouns. Plural marking otherwise seems to be the norm.

A.61 *Taleshi* [*taly1247*] (*+clf,-obl.pl*)
Taleshi has a number of numeral classifiers, use of which is non-obligatory (Paul 2011:181–182); additionally, "any noun following a numeral phrase is generally in the singular." Elsewhere, plural number is marked with a suffix that varies from dialect to dialect.

A.62 *Torwali* [*torw1241*] (*-clf,morph.pl*)
According to Grierson (1929:34), if a noun ends in a vowel, it can take a plural-marking suffix *-e*; otherwise, the singular and plural forms are identical.

A.63 *Wakhi* [*wakh1245*] (*+clf,-obl.pl*)
Native Wakhi numeral forms are in competition with Tajik numeral forms (Grjunberg & Steblin-Kamenskij 1988:89–90). The word for twenty (*bist*) is borrowed from Tajik, but numerals 20–30 can combine *bist* with either Wakhi or Tajik forms in the digits place. A number of classifiers are borrowed from Tajik. Unlike the situation in Yaghnobi, these classifiers can be used with both Wakhi and borrowed Tajik numerals. Plural is marked with the suffix *-iš*(*t*) (p. 19), but this marking appears to be optional.

A.64 *Yaghnobi* [*yagn1238*] (*+clf,+obl.pl*)
According to Xromov (1972:21–22), when numbers two and upward combine with nouns, the noun is found in the oblique singular form; if measure terms are used, the measure term is marked for oblique singular. The numerative *ta*, borrowed from Tajik, can be used, but only with Tajik numerals. Outside of the context of enumeration, *-t/d* is consistently used to mark plural number on nouns.

A.65 *Zazaki* [*diml1238*] (*+clf,-obl.pl*)
According to Paul (1998:19 ff.), a morphologically singular noun can be used in a generic sense, but nouns denoting a plurality, definite or indefinite, take the plural ending. Plural marking is non-obligatory. An apparent sortal classifier *teney* co-occurs with nouns marked both for singular and plural.